DOCUMENT RESUME

ED 146 195 TM 006 569

AUTHOR Mintzes, Joel J.
TITLE Field Test and Validation of a Teaching Evaluation Instrument: The Student Opinion Survey of Teaching. A Report Submitted to the Senate Committee for Teaching and Learning, Faculty Senate, University of Windsor, Windsor, Ontario.
INSTITUTION Windsor Univ. (Ontario).
PUB DATE [77]
NOTE 97p.; Pages 85 through 119 of the original document are copyrighted and therefore not available. They are not included in the pagination

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS Academic Achievement; Class Size; *College Students; College Teachers; *Course Evaluation; Higher Education; Predictor Variables; *Questionnaires; Student Characteristics; *Student Evaluation of Teacher Performance; Teacher Characteristics; *Test Reliability; *Test Validity
IDENTIFIERS *Student Opinion Survey of Teaching

ABSTRACT

The reliability and validity of the Student Opinion Survey of Teaching (SOST) were assessed, and normative data were provided for judging its value in the evaluation of faculty teaching. Data were collected from 2,229 students enrolled in 93 classes taught by 53 instructors in 12 academic disciplines at the University of Windsor, Ontario. Internal consistency of the SOST was moderate to relatively high on three sections; however alpha coefficients for two sections were unacceptably low. Low but significant positive correlations were found between eleven of the SOST items and student achievement in an introductory psychology course. These findings provided evidence that the instrument possesses a certain degree of criterion-related validity. Five factors accounted for 55% of the variance in item responses: instructional skill, student teacher interaction, work load, instructor's organization of the course, and feedback. Results indicated that a student's major may affect his or her evaluations of courses and instructors, upper-level students tend to rate instructors more favorably than lower-level students, superior or above average students tend to give their instructors better ratings, and elective courses are rated more favorably than required courses. The author suggests that although the SOST is valid and reasonably stable, the instrument should not be adopted in its present form without revision and further testing (HV)

FIELD TEST AND VALIDATION OF A TEACHING EVALUATION INSTRUMENT: THE STUDENT OPINION SURVEY OF TEACHING[1]

A REPORT
SUBMITTED TO THE SENATE COMMITTEE
FOR TEACHING AND LEARNING
FACULTY SENATE
UNIVERSITY OF WINDSOR
WINDSOR, ONTARIO

Joel J. Mintzes

JOEL J. MINTZES
ASSISTANT PROFESSOR
DEPARTMENT OF BIOLOGY

With the Assistance and Cooperation of:

LAUREL BROWN
DORIS COMPTON
DAVID V. REYNOLDS
DEPARTMENT OF PSYCHOLOGY

UNIVERSITY OF WINDSOR
WINDSOR, ONTARIO
1976-77

[1]This research was supported by a grant from the Ontario Universities Program for Instructional Development (QUPID) through the University of Windsor Office of Learning-Teaching Development (Professor W. Romanow, Coordinator).

TABLE OF CONTENTS

LIST OF TABLES

I. INTRODUCTION

1.1 OBJECTIVES

The objectives of this study were to assess the reliability and validity of the Student Opinion Survey of Teaching (SOST) and to provide normative data for judging the usefulness of this instrument in the evaluation of faculty teaching at the University of Windsor.

1.2 BACKGROUND

At the January 25, 1975 meeting of the University of Windsor Faculty Senate, a resolution was passed to establish a special committee "...to review the present practices and procedures for student evaluations of teaching performance." This committee presented an interim report in December of 1975 (Student Evaluations Committee).

In addition to proposing a University policy on teaching evaluation, the Senate Student Evaluations Committee devoted a considerable amount of time and energy to the development of a survey instrument for eliciting student opinion of teaching performance. In developing the Student Opinion Survey of Teaching (SOST), the Committee examined and analyzed a large number of similar questionnaires used by both Canadian and American universities.

In its December report, the Committee recommended:

(a) that the SOST be adopted for University-wide evaluations by all Faculties and Departments;

(b) that a student committee be charged with coordinating survey activities, validating and updating the instrument, and interpreting the data;

(c) that the results of faculty evaluations be made available to instructors, students, promotion and tenure committees, and University administration.

1.3 THE PROBLEM

Following the December report, concern was expressed by a number of individuals that the immediate adoption and university-wide dissemination of SOST results would be inappropriate and even negligent until

the instrument had been field-tested and, if necessary, revised. It was felt that the gathering and dissemination of data which could directly affect the promotion and tenure of large numbers of faculty members should be based on an instrument of known reliability and validity. Furthermore, some felt the field-test and validation should be conducted by unbiased individuals who had no part in the development of the instrument.

The present study, supported by funds from the Ontario Universities Programme for Instructional Development through the Office of Learning-Teaching Development, examined the reliability and validity of the SOST (Appendix A). Recommendations for revision are included in Section IV of this report.

1.4 EXPERIMENTAL DESIGN

This section describes the general experimental design employed in the study. Detailed Descriptions of subjects, data collection, and analytic procedures are included in Section III.

1.41 Subjects

The total data pool represents the responses of 2229 students who were enrolled in 93 classes taught by 53 instructors in 12 academic disciplines. Although an attempt was made to sample student opinion from a wide range of subject areas, the data do not necessarily represent a random cross-campus sample of students, courses or instructors.

Participation in the study by instructors was voluntary. In most cases individual instructors were contacted verbally and consent was obtained by a follow-up explanatory letter (Appendix B). In a few cases department heads and deans were asked to approach individual faculty members on a voluntary basis.

Table 1.1 sumarizes the data pool, indicating the number of student responses, instructors and classes by subject area. Approximately one-half of the responses were obtained among students enrolled in Biology and Psychology courses. These two departments also accounted for over 50% of the instructors (30/53) and classes (61/93).

Table 1.2 presents a profile of instructors by rank and sex. The category "other", which includes, Lecturers, Instructors and Teaching Assistants, accounted for almost 50% of the instructors by rank. The fewest ratings were obtained among Full Professors (10%). Approximately 60% of the instructors were males; 40% were females.

Table 1.1 SUMMARY OF DATA BY SUBJECT AREA

Subject Area	Instructors	Classes	Student Responses
Biology	13	36	670
Business Administration	1	2	97
Chemistry	1	1	79
Education	2	2	32
Engineering	1	1	9
Geology	1	2	59
Germanic & Slavic Studies	4	9	106
Mathematics	1	2	105
Nursing	8	9	265
Philosophy	2	2	29
Psychology	17	25	622
Sociology & Anthropology	2	2	156
TOTALS	53	93	2229

Table 1.2 PROFILE OF INSTRUCTORS BY RANK AND SEX

Rank	Sex		Total
	Male	Female	
Professor	4	1	5
Associate Professor	9	2	11
Assistant Professor	8	4	12
Other	10	15	25
TOTALS	31	22	53

A profile of the student raters is given in Table 1.3. About 65% of the students in the sample were Science and Mathematics or Social Science Majors. These students were fairly evenly distributed between Honours and General programmes. Over 50% of the students were enrolled in their first year of university.

1.42 Data Collection

Two part-time research assistants were employed to help in the data collection and organization. The general procedure was as follows:

a) the assistant arrived at the agreed-upon (instructor-selected) time and the instructor was asked to leave the room.

b) the students were asked to cooperate in the evaluation of the instructor but were not informed that responses would be used for research purposes.

c) each student received a copy of the instrument (SOST) and a multiple choice standard response form for recording his/her evaluations.

d) students were asked to indicate the course number and the instructor's name on the response form but to omit their own name and student number.

e) depending on class size, the students were permitted 10 to 15 minutes to complete the evaluation. The response forms and instruments were then collected and the students were thanked for their cooperation.

f) the completed response forms were optically scanned and the data transferred to standard data processing cards.

All evaluations were completed during the last 4 weeks of the Fall and Spring terms, 1976-77. Table 1.4 presents a summary of the data, giving means and standard deviations for each of the evaluative items (9-28). This summary is based on the entire data pool (2229 responses) regardless of class or instructor.

1.43 Analyses

All analyses were performed with the aid of the University IBM

Table 1.3 PROFILE OF STUDENT RATERS
(N=2229)

Item	Frequencies				
	A	B	C	D	E
1. My major is in:	Arts	Soc. Sci.	Sci. & Math	Bus.	Other
	13.1%	24.6%	40.3%	8.9%	13.1%
2. This course is part of my:	Hon. Pgm	Gen. Pgm.			
	52.1%	47.9%			
3. I have completed the following number of University-level full courses:	0-2	3-7	8-12	13-17	18--
	52.0%	15.4%	10.5%	8.5%	13.7%
4. Rating myself against the performance of other students in the class, I see myself in one of the following groups:	Superior	Above Avg.	Average	Below Avg.	Failing
	4.9%	38.6%	49.5%	6.0%	1.0%
5. This course was compulsory.	Yes	No	Not Sure		
	51.7%	44.2%	4.1%		
6. My Attendance and Punctuality have been consistently good.	Yes	No			
	91.1%	8.9%			
7. Compared to other courses I have taken, I consider my effort in this course to have been:	Excellent	Above Avg.	Average	Below Avg.	Poor
	10.3%	39.9%	41.6%	6.9%	1.2%
8. I have found the material in this course to be inherently difficult	Yes	No			
	29.3%	70.7%			

Table 1.4 SUMMARY OF SOST-DATA: MEANS & STANDARD DEVIATIONS[1,2]

ITEM	MEAN	STANDARD DEVIATION
9	1.900	0.912
10	2.244	1.060
11	3.756	1.014
12	1.869	0.855
13	1.949	0.965
14	1.461	0.668
15	1.848	0.902
16	3.627	0.936
17	2.009	0.873
18	2.353	1.076
19	3.236	1.147
20	2.778	1.022
21	2.464	0.976
22	2.654	0.917
23	2.453	1.016
24	3.653	0.988
25	3.513	0.760
26	3.518	0.966
27	2.808	1.227
28	2.436	0.976

1 Based on 2229 student responses in 93 class sections taught by 53 instructors in 12 academic disciplines.

2 Responses were coded as follows: A = 1, B = 2, C = 3, D = 4, and E = 5.

360/65 computing facility using subprograms of the Statistical Package for the Social Sciences (SPSS) and the Statistical Analysis System (SAS). The following is a general outline of the analyses:

Normative Data (Appendix C). These are descriptive data including means and standard deviations for each of the items on the instrument. These data differ from those in Table 1.4 in that all analyses were based on class means. These data permit individual instructors to compare their own class evaluations with "average" ratings obtained in 93 other classes.

Reliability. Both internal consistency and stability were examined. The internal consistency of each subscale was estimated using Cronbach's alpha coefficient. Stability was assessed by the test-retest procedure with intervals of 7, 14, 21, and 28 days.

Validity. The construct validity of the instrument was examined by factor analysis. In addition, a series of analyses of variance were performed to determine whether student responses were systematically biased by irrelevant factors (student characteristics; instructor characteristics; class characteristics). Criterion-related validity was assessed by examining relationships between student ratings and objective measures of student achievement. Finally, correlations were obtained between each of the SOST items and each item on a number of other widely-used teaching evaluation instruments.

1.5 ORGANIZATION OF REPORT

Section II of this report presents a brief review of representative studies on the reliability and validity of teaching evaluation instruments. Section III reports the reliability and validity of the SOST and Section IV gives a summary of the findings, presents the conclusions, and forwards several recommendations concerning the development and use of teaching evaluation instruments at the University of Windsor.

II. RELIABILITY AND VALIDITY OF TEACHING EVALUATION INSTRUMENTS

2.1 RELIABILITY

Two terms that are often used to describe the meaning of reliability are "precision" and "consistency." In test and measurement theory reliability is an estimate of the extent to which "differences in (observed) test scores are attributable to 'true' differences in the characteristics under consideration and the extent to which they are attributable to chance errors (Anastasi, 1976). Said another way, the reliability of a test is a measure which allows us to estimate what proportion of observed test score variance is error variance.

Gulliksen (1965) defines these relationships concisely as follows:

$X_i = T_i$ & E_i or $E_i = X_i - T_i$

Where: X_i = the observed score of the i^{th} person

T_i = the "true" score of the i^{th} person

E_i = the error component for the same person

It is apparent then, that the observed score for any individual is a composite of that individual's "true" score and an error factor. Furthermore, the variance of observed scores in a population (S^2_X) is made up of "true" score variance (S^2_T) and error variance (S^2_E), and the reliability coefficient is the ratio:

$$r_{XX} = \frac{S^2_T}{S^2_X}$$

An extremely important question is, "what factors add to the error variance thereby affecting the reliability of a given test?" To help answer this question Thorndike (1949) and Cronbach (1970) have classified the sources of test score variance (Table 2.1). The sources of variance include: (1) lasting - general characteristics such as reading and problem solving abilities, (2) lasting - specific characteristics such as knowledge of specific test questions, (3) temporary - general characteristics such as health, fatigue and motivation,

Table 2.1 SOURCES OF TEST-SCORE VARIANCE CLASSIFIED*

I. Lasting and general characteristics of the individual.
   1. General skills (e.g., reading)
   2. General ability to comprehend instructions, testwiseness, techniques of taking tests
   3. Ability to solve problems of the general type presented in this test
   4. Attitudes, emotional reactions, or habits generally operating in situations like the test situation (e.g., self-confidence)

II. Lasting and specific characteristics of the individual.
   1. Knowledge and skills required by particular problems in the test
   2. Attitudes, emotional reactions, or habits related to particular test stimuli (e.g., fear of high places brought to mind by an inquiry about such fears on a personality test)

III. Temporary and general characteristics of the individual (systematically affecting performance on various tests at a particular time)
   1. Health, fatigue, and emotional strain
   2. Motivation, rapport with examiner
   3. Effects of heat, light, ventilation, etc.
   4. Level of practice on skills required by tests of this type
   5. Present attitudes, emotional reactions, or strength of habits (insofar as these are departures from the person's average or lasting characteristics — e.g., political attitudes during an election campaign)

IV. Temporary and specific characteristics of the individual.
   1. Changes in fatigue or motivation developed by this particular test (e.g., discouragement resulting from failure on a particular item)
   2. Fluctuations in attention, coordination, or standards of judgment
   3. Fluctuations in memory for particular facts
   4. Level of practice on skills or knowledge required by this particular test (e.g., effects of special coaching)
   5. Temporary emotional states, strength of habits, etc., related to particular test stimuli (e.g., a question calls to mind a recent bad dream)
   6. Luck in the selection of answers by "guessing"

*After R.L. Thorndike, 1949, p.73 and L.J. Cronbach, 1970, p. 175.

and (4) temporary - specific characteristics such as fluctuations in attention, coordination and memory for specific facts.

The lasting-general characteristics affect the "true" score variance and the temporary-specific characteristics affect the error variance. The lasting-specific characteristics and temporary-general characteristics may affect either variances depending upon the type of reliability being studied.

Research on teaching evaluation instruments has concentrated on two types of reliability: internal consistency and stability. In general, internal consistency studies examine the degree of homogeneity of items and/or behaviours sampled by a test or subscale. Tests of internal consistency such as Cronbach's alpha (Cronbach, 1970) count lasting-specific and temporary-specific characteristics as error variance. On the other hand, studies of stability provide an index of the extent of fluctuation in scores over a specified time interval. The test-retest procedure, a commonly employed measure of stability, considers temporary-general and temporary-specific characteristics as sources of error variance.

Internal consistency and stability are independent of each other. An internally consistent instrument may or may not be stable. A stable instrument may or may not be internally consistent.

2.11 Internal Consistency Studies

A large number of studies on the internal consistency of student evaluation instruments have been reported over the past 25 years. Table 2.2 provides a fairly representative sample of these studies. We will not attempt here to evaluate or even discuss each of these studies individually, however, a number of cautionary statements seem appropriate.

Even a casual examination of the research shows a great deal of variability among studies with regard to:

1) numbers of student raters (59 to 35,000) and numbers of courses (1 to 1279),
2) numbers of items per subscale or instrument (2 to 140), and
3) analytic procedures used (split-half, Kuder-Richardson formulae, odd-even means, Cronbach's alpha, Hoyt and others). Each of these differences may affect the interpretation of the internal consistency coefficient.

Perhaps the most significant of the above mentioned sources of variability is the number of items per subscale. It has long been recognized (Spearman, 1910 and Brown, 1910) that, other things being equal, the longer a test the more reliable (internally consistent) it is. Therefore, care should be taken when comparing internal consistency scores across subscales possessing different numbers of items.

Another source of variability among the studies is the unit of analysis employed. In some studies the unit of analysis is individual students within a class (example: Wherry, 1951). In other studies it is students across classes (example: Aleamoni and Spencer, 1973). In still other studies, the analytic unit is class means presumably regardless of individual class size (Pohlmann, 1975). Here again, because of differences in experimental design, caution should be exerted when comparing coefficients among studies.

With these cautions in mind, one might still be impressed with the remarkably high internal consistency coefficients reported. The coefficients compare quite favorably with many psychometric instruments including ability tests and personality inventories, even those developed by factor analytic techniques.

Table 2.2 REPRESENTATIVE STUDIES OF INTERNAL CONSISTENCY

Researcher	Student Raters (N)	Type of Instrument	Analytic Procedure	Internal Consistency
Wherry (1951)	42 rating past better and worse instructors	12-item 25-point ratings	Split-half	.88
	46	140-item 25-point ratings	Split-half	.98
	46	140-item five-point ratings	Split-half	.96
	47	12-item five-point ratings	Split-half	.98
	44	70 forced-choice dyads	Kuder-Richardson 14	.88
Lovell and Haner (1955)	105 in 4 courses	36 forced-choice tetrads	Odd-even means	.79 corrected to .88
Remmers and Weisbrodt (1965)	1908 in 59 courses	11 ten-point ratings (PRSI)	Horst	.67-.91
Harvey and Barker (1970)	59 male students regardless of courses	21-item ten-point ratings	Product-moment correlation	.38-93
Hildebrand, Wilson, and Dienst (1971)	1015 rating past best and worst instructors	7-8 item seven-point ratings	Alpha	.80-.89
Doyle (1972)	379 in 11 courses	9-28 item 5-point ratings (SOS[c])	Hoyt	.90-.96
Aleamoni and Spencer (1973)	297 regardless of courses	50-five-point ratings (CEQ[b])	Split-half (negative versus positive items)	.85 corrected to .92
		50-five-point ratings	Split-half (mixed negative and positive)	.87 corrected to .93

...continued

Table 2.2 REPRESENTATIVE STUDIES OF INTERNAL CONSISTENCY*

Researcher	Student Raters (N)	Type of Instrument	Analytic Procedure	Internal Consistency
	? in 16 courses	50-five-point ratings	Kuder-Richardson 21	.93 average
	94-571 in 7 courses	8-10 item CEQ[b] subscales	Kuder-Richardson 21	.40-.92
Pohlmann (1975)	35,000 in 1,279 courses	21-item 5-point ratings	Product-moment correlation	.52-.87

[a]Purdue Rating Scale for Instructors

[b]Illinois Course Evaluation Questionnaire

[c]Minnesota Student Opinion Survey

*Modified and Updated after Doyle (1975)

It seems then that many widely-used teaching evaluation instruments do possess relatively high internal consistency with coefficients averaging approximately .7 to .9.

2.12 Stability Studies

Research on the stability of student evaluation instruments has extended over a period of 50 years beginning with the work of Remmers and his associates (1927) on the Purdue Rating Scale for Instructors. Although a considerable number of studies have been reported, the usefulness of this kind of information has been questioned by several educators and psychometricians. Says Doyle (1975):

> While it would be important to know the extent to which ratings change over time as a function of random or systematic rater, task, and situational factors as distinguished from instructor and course factors, the typical retest study is only marginally adequate to the task, given that instructor changes are uncontrolled and trait differences usually unexamined.

Researchers conducting stability studies have responded that teaching behaviours tend to remain stable over short periods of time even, for example, when instructors are given feedback by way of student evaluations (Murray, 1973). Therefore, retest studies are helpful in identifying the extent of a major source of error variance that attributable to fluctuations in student characteristics.

As with studies of internal consistency, the design of stability studies varies considerably. The major differences among studies are: number of student raters, type and number of items, and very importantly, the time interval between initial test and retest. (In general, the stability of an instrument decreases with increasing time intervals (Anastasi, 1976)).

All of the studies summarized in Table 2.3 with the exception of Bausell et. al. (1975) examined stability within individual courses over relatively short periods of time (3 days to one semester). It is probable that the majority of the variance in stability coefficients among these studies is attributable to differences in the items and

numbers of student raters (critical r values decrease with increasing N). In general, these studies indicate that student evaluation instruments possess moderate to high stability over short time intervals.

Another objection to stability studies is that voiced by Kulik and McKeachie (1975):

> For educational administrators and researchers, the meaning of these reliability coefficients is limited. Reliability coefficients for individual student ratings reflect the degree of consistency of students. Most educational administrators and researchers are concerned with the consistency of teachers and therefore are more concerned with the reliability of class ratings of instructors.

The study by Bausell et. al. (1975) is particularly interesting in that it addresses the problems posed by Kulik and McKeachie. Rather than using individual student ratings, class means were computed. Class means for "same course - same instructor" were correlated across time intervals ranging from one semester to two years. In essence, then, this study examined the combined stability of teaching behaviours and student ratings. The data are confounded, however, because raters changed with class enrollment.

Nonetheless, several interesting conclusions might be drawn by comparing Bausell's mean stability coefficients (.64, .78 and .65) with those from within individual courses (Table 2.3). It appears that the two are not strikingly dissimilar. This would argue that the majority of the variability in student ratings can be attributed to the rater and that teaching behaviours remain fairly stable even over longer periods of time.

Costin, Greenough and Menges (1971) summarize their review of reliability studies in the following way:

> It would appear, then, that students can rate classroom instruction with a reasonable degree of reliability. In particular, the evidence cited concerning the stability of students' ratings argues against the contention ..... that student opinions of instruction are difficult to interpret since they might be made after a particularly good or bad atypical experience (e.g., a lecture).

Table 2.3 REPRESENTATIVE STUDIES OF STABILITY*

Researcher	Student Raters (N)	Type of Instrument	Time Interval	Stability
Remmers and Brandenburg (1927)	30-33 in 3 courses	10-ten-point ratings (PRSI[a])	3 days	.42-.92
Root (1931)	200 in one course	50 item checklist	4 weeks	.95
Lovell and Haner (1955)	105 in 4 courses	36 force-choice tetrads	2 weeks	.89
Costin (1968)	Unreported number, mostly in sections of one large course	5 subscales, 3-5 five-point ratings each	Mid-semester to end of semester	.41-.87
Kooker (1968)	92 in 4 sections	7 subscales, 7-14 five-point ratings each	2 weeks	.58-.87
	Same	Total scores	2 weeks	.91
Costin (1971)	219 of 11 instructors	4 subscales, 4-7 five-point ratings each (factor scores)	2 weeks	.67-.77
Kohlan (1973)	271 in eight classes	3 subscales, 4-7 five-point ratings each	2nd day of semester to last week of semester	.55-.70
		Single general five-point ratings	2nd day of semester to last week of semester	.58
Bausell et al (1975)	41 courses	14 5-point ratings	Fall 1968 to Spring 1969	.31-.79 ($\bar{x}$=.64)
	39 courses	12 5-point ratings	Fall 1972 to Fall 1973	.64-.87 ($\bar{x}$=.78)
	37 courses	10 5-point ratings	Spring 1973 to Fall 1973	.23-.80 ($\bar{x}$=.65)

[a]Purdue Rating Scale for Instructors *Modified and Updated after Doyle (1975)

2.2 VALIDITY

In general, the validity of a psychometric instrument is an assessment of what an instrument measures and how well it measures it. However, these criteria are not sufficiently specific to convey the entire meaning of validity.

According to Standards for Educational and Psychological Tests and Manuals (1974), a joint effort of the American Psychological Association, American Educational Research Association and the National Council on Measurement in Education, "validity information indicates the degree to which the test is capable of achieving certain aims." The aims of testing may be:

1. .... to determine how an individual performs at present in a universe of situations that the test situation is claimed to represent. (ex: school achievement tests)
2. .... to forecast an individual's future standing or to estimate an individual's present standing on some variable of particular significance that is different from the test. (ex: scholastic aptitude tests)
3. .... to infer the degree to which the individual possesses some hypothetical trait or quality (construct) presumed to be reflected in the test performance. (ex: personality tests)

Based on these "aims of testing" three types of validity have been recognized: content validity, criterion-related validity, and construct validity.

Content validity is a measure of the extent to which the instrument samples the universe of behaviours about which generalizations are to be made. This type of validity is important in scholastic and vocational tests of knowledge or specific skills. For example, a valid test of arithmetic ability would contain a representative sample of addition, subtraction, multiplication and division problems at varying levels of difficulty and abstraction.

In order to examine the content validity of a test one must first do a systematic analysis of the behaviour domain covered by

the test. Subsequently, one would examine the number and levels of specific test items to insure proportional representation and completeness of coverage.

Criterion-related validity is a measure of how well a test predicts an individual's behaviour in a given set of circumstances. Usually it is demonstrated by correlating test scores with some independent measure of performance. For example, the Medical College Admissions Test (MCAT) is designed to predict academic success in medical school. To the extent that it does so with some degree of accuracy, it may be said to have criterion-related validity.

The APA Standards identifies two types of criterion-related validity which differ in the time interval between test administration and criterion measurement. If the two are separated by a reasonable period of time, the measure of association between them is referred to as "predictive" validity. This kind of information is most useful in making decisions such as personnel hiring or classification and student admissions or placement.

If a test is administered to an individual or group of individuals on whom criterion data is already available, the relationship is referred to as "concurrent" validity. Knowledge of concurrent validity is helpful in interpreting the results of tests which examine the present status of an individual (patient, student, employee) rather than predicting future status.

Construct validity is a measure of the degree to which test scores reflect some hypothetical or theoretical trait or factor. Examples of such factors (constructs) are: "school motivation", "social introversion", "manual dexterity" and "mathematical aptitude".

Construct validity is especially important in the clinical application of personality inventories for diagnosing behavioural and emotional disorders. A number of these tests, including the Minnesota Multiphasic Personality Inventory (MMPI) and the California Psychological Inventory (CPI), are purported to have a high degree of construct validity.

Several procedures are commonly used to assess the construct validity of a particular test. The most important of these are correlations

with other tests which measure the same construct, and factor analysis.

Factor analysis is a fairly sophisticated analytic procedure that is often useful in identifying commonalities in behavioural data. The goal of factor analysis is to see whether some underlying pattern of relationships exists such that the data may be re-arranged or reduced to a smaller set of factors. This is accomplished by analyzing all possible correlations among items, grouping common items together, and assigning appropriate loadings (weights) to each. The clusters so formed may represent unique factors or constructs.

2.21 Validity Studies: Overview

The validation of a teaching evaluation instrument is an especially difficult task when compared to the validation of other test and measurement devices. Each type of validity poses special problems.

Content Validity. Presumably a teaching evaluation instrument possessing content validity is one that has a representative number (and type) of items which measure teaching behaviours that affect student learning. Unfortunately, the plain truth of the matter is that we know very little about this domain of behaviours. In fact, it has been suggested by some (Butts and Capie, 1977) that teaching behaviours may account for less than 10% of the variance in student achievement.

Another approach is to ask students the criteria that they consider important to teaching effectiveness and then to examine teacher rating forms for completeness based on these criteria. In fact, this may be a better measure of content validity in that these instruments are designed to measure the student's perception of teaching effectiveness (Table 2.4).

The best and most widely used teaching evaluation instruments are assumed to possess content validity because, as a first step in their construction, students (and faculty) are often asked to list characteristics of particularly good and/or particularly poor instructors (Hildebrand et. al, 1971). The list of characteristics is then trimmed by data reduction procedures (often factor analysis) and phrased into items to produce the final instrument.

Interestingly, the factors most often named as characteristic of

Table 2.4 CHARACTERISTICS OF GOOD TEACHING*

Bousfield[1]	Clinton[2]	Deshpande, et al[3]	French[4]
Fairness	Knowledge of subject matter	Motivation	Interprets ideas clearly
Mastery of subject	Pleasing personality	Rapport	Develops student interest
Interesting presentation of material	Neatness in appearance and work	Structure	Develops skills of thinking
Well-organized material	Fairness	Clarity	Broadens interests
Clearness of exposition	Kind and sympathetic	Content mastery	Stresses important materials
Interest in students	Keen sense of humor	Overload (too much work)	Good pedagogical methods
Helpfulness	Interest in profession	Evaluation procedure	Motivates to do best work
Ability to direct discussion	Interesting presentation	Use of teaching aids	Knowledge of subject
Sincerity	Alertness and broadmindedness	Instructional skills	Conveys new viewpoints
Keenness of intellect	Knowledge of methods	Teaching styles	Clear explanations
1. Listed in order of importance, by 61 undergraduates at University of Connecticut.	2. Listed in order of importance, by 177 junior-years students at Oregon State University.	3. Listed in order of importance, by 674 undergraduates who rated 32 engineering teachers.	4. Listed in order of importance, by undergraduates at the University of Washington

*After R.I. Miller, 1974.

.....cont'd

Table 2.4 CHARACTERISTICS OF GOOD TEACHING ...con'td

Gadzella[5]	Perry[6]	Pogue[7]	Hildebrand[8]
Knowledge of subject	Well-prepared for class	Knowledge of subject	Dynamic and energetic person
Interest in subject	Sincere interest in subject	Fair evaluator	Explains clearly
Flexibility	Knowledge of subject	Explains clearly	Interesting presentation
Well-prepared	Effective teaching methods		Enjoys teaching
Uses appropriate vocabulary	Tests for understanding		Interest in students
	Fair in evaluation		Friendly toward students
	Effective communication		Encourages class discussion
	Encourages independent thought		Discusses other points of view
	Course organized logically		
	Motivates students		

5. Listed in order of importance, by 443 undergraduates at Western Washington State College

6. Listed in order of importance, 1493 students, faculty, alumni at University of Toledo.

7. Listed in order of importance, 307 students at Philander Smith College.

8. Listed in order of importance, by 338 undergraduate and graduate students at University of California, Davis.

good instructors seems to be fairly constant. Thus many instruments have items covering the following areas:

a) knowledge of subject
b) ability to present material in a clear and interesting manner
c) ability to motivate students
d) course organization
e) course workload
f) rapport
g) feedback
h) student evaluation (grading procedures)

Criterion-related Validity. The problem with criterion-related validity is that there is a general lack of agreement on the criteria. Most people, however, would concede that a "good teacher" is one who facilitates student learning. As a result, many studies have examined relationships between instructor ratings and student achievement as the best estimate of criterion-related validity. The major problem with this, of course, is that many factors may affect learning; teaching behaviours constitute only one of these factors (possibly one of the least significant factors). These studies, then, are confounded by many variables not directly related to the teaching behaviours of the instructor.

Another approach has been to examine relationships between student ratings of instructors and ratings of the same instructors given by other individuals. In most cases the "other individuals" are colleagues, department chairmen, deans, alumni or paid observers. In at least one study (Centra, 1973) an attempt was made to find relationships between student ratings and self-ratings by instructors. In any case, mere agreement between students and other observers probably constitutes a weak measure of validity.

Construct Validity. The major problem with construct validity is that "constructs" concerning teaching effectiveness are rather vague and ill-defined. For example, the construct, "ability to motivate students", may mean several things to different people or even several things to the same individual.

None-the-less, two approaches have generally been used to assess the construct validity of teacher rating forms. One method is to examine correlations between instruments. If two instruments measure the same constructs, corresponding items on the instruments should be highly correlated. The second approach, and one which has met with a reasonable amount of success, is factor analysis. If an instrument possesses construct validity, subscale factor loadings should be relatively high.

Perhaps the most common complaint voiced by opponents of teaching evaluations is that they are subject to student bias and that ratings are affected by many variables over which the instructor has no control. To test these hypotheses studies have examined the effects of a number of student, instructor, and class variables on teacher ratings.

The remainder of this section summarizes the results of representative studies on:

1) Criterion-related Validity: Student Ratings and Ratings of Others
2) Student Ratings and Achievement
3) Construct Validity: Factor Analysis
4) Effect of Student Variables on Ratings
5) Effect of Instructor Variables on Ratings
6) Effect of Class Variables on Ratings

2.22 Student Ratings and Ratings of Others

Several studies have examined correlations between student ratings and ratings given by colleagues (Table 2.5). In general, the coefficients obtained have been moderate, averaging .4 to .5. One study (Murray, 1973) found a correlation of .82.

Ratings given to teaching assistants in large courses by their supervisors also correlate moderately well with student evaluations (.49; .62) as do alumni ratings of their former professors (~.5).

The highest degree of agreement has been found between ratings given by paid observers and student ratings (.92). Perhaps this study

Table 2.5 STUDENT RATINGS AND RATINGS OF OTHERS

Study		Correlation Coefficients
I. Colleague Ratings		
Maslow and Zimmerman	(1956)	.30 to .63
Algamoni and Yimer	(1973)	.16 to .30
Murray	(1973)	.82
Centra	(1975)	.00 to .54
II. Supervisor Ratings		
Hayes	(1971)	.62
Costin	(1966)	.49
III. Alumni Ratings		
Drucker and Remmers	(1950)	.40 to .68
IV. Paid Observer Ratings		
Murray	(1973)	.92
V. Self-ratings		
Centra	(1973)	.21

is the most significant in that the paid observers attended classes with students and were therefore in a better position than other outside raters to judge the effectiveness of a given instructor.

Finally, one study (Centra, 1973) examined relationships between student ratings and instructors' self-ratings. The correlations obtained were low, averaging .21. In most cases instructors tended to rate themselves more favourably than their students.

2.23 Student Ratings and Achievement

Another way of looking at criterion-related validity is to examine relationships between ratings and student achievement. According to Murray (1973),

> Students and faculty would agree that the ultimate criterion of good teaching is the extent to which students learn, or make progress toward educational goals. Most rating forms for student evaluation of teaching are not intended to provide a direct measure of student learning, but they are designed to measure aspects of teaching (eg. clarity of presentation) that would be expected to have some direct or indirect effect upon student learning. Thus it is reasonable to expect some degree of positive correlation between student ratings of teaching and objective measures of student achievement.

Many studies have examined the correlation between ratings and student achievement (Table 2.6). In general, the evidence shows a weak but positive relationship with correlation coefficients averaging about .20 to .30. This indicates that instructors who obtain favourable ratings are more effective in facilitating learning than instructors who receive less favourable ratings.

One study by Rodin and Rodin (1972) has received considerable attention and caused a certain amount of controversy. The findings of this study show a strong negative relationship (-.75) between ratings and achievement. One of the reasons it has caused so much discussion is that the findings were reported in Science, the prestigious journal of the American Association for the Advancement of Science. Nevertheless, the results have been severely criticized on methodological grounds by several individuals including Frey (1973) whose parallel study (also

Table 2.6 STUDENT RATINGS AND ACHIEVEMENT

Study	Correlation Between Ratings and Achievement
Elliott (1950)	+ .24
Morsh et. al. (1956)	+ .40
McKeachie et. al. (1971)	-.60 to + .72 ($\bar{x}$= + .10)
Rodin and Rodin (1972)	- .75
Frey (1973)	+ .14 to + .91
Gessner (1973)	+ .53
Skanes and Sullivan (1974)	+ .39
Marsh et. al. (1975)	- .02 to + .55

published by Science) shows equally strong but positive correlations.

2.24 Construct Validity: Factor Analysis

A large number of factor analytic studies have been performed on student ratings. The studies have generally been of two types. One type of study attempts to extract the overall dimensions or factors describing "good teaching" from pools of statements submitted by students and faculty. This type of study has often been performed preliminary to the development of a new teaching evaluation instrument. A second type of study factor analyzes an existing instrument to determine its factor structure for practical use in a college or university setting.

Probably the most influential factor analytic study was performed by Isaacson, McKeachie, Milholland et. al. (1964). In this study a pool of 145 items describing teachers was reduced to 46 representative statements. The 46 items were then factor analyzed for four separate student samples. Six factors emerged and were consistently found with two administrations, in different semesters with different students and teachers. The 6 factors were labelled "Skill", "Rapport", "Structure", "Overload", "Feedback", and "Evaluation."

The first four of these items seem to correspond to similar factors emerging from 11 other studies (Table 2.7). It should be noted however that factor labels are derived somewhat subjectively and therefore similarities may be misleading. It is remarkable though to observe the amount of agreement in studies spanning a period of 30 years. It would seem that our basic conception of what constitutes "good teaching" has not been altered significantly even with the introduction of new instructional methods and modern technological advances.

2.25 Effect of Student Variables on Ratings

Over the past 50 years many student variables have been examined as possible sources of bias in student ratings. While a number of studies have looked at personality characteristics, the factors most often studied have been demographic in nature including students' sex, major, level (year in university), and course grades (Table 2.8).

The results of these studies have been quite variable because of differences in experimental design and methodological rigour. Nevertheless,

Table 2.7 FACTOR ANALYSES OF STUDENT RATING INSTRUMENTS[1]

Study	Skill	Rapport	Structure	Overload	Other
Smalzreid & Remmers (1943)	Professional Maturity	Empathy			
Creager (1950)	Professional Impression	Rapport			
Bending (1954)	Instructor Competence	Instructor Empathy			+ one other factor
Gibb (1955)	Communication	Friendly-Democratic	Organization	Academic Emphasis	
Isaacson et. al. (1964)	Skill	Rapport	Structure	Overload	+ two other factors
Solomon (1966)	Energy vs. Lethargy	Lecturing vs. Student Participation	Control vs. Permissiveness		
Turner (1970)	Exciting, Humorous, Stimulating	Approachable, Warm, Cheerful	Penetrating, Clear, Focused	Prepared, Probing, Demanding	+ two other factors
Deshpande et. al. (1970) 2nd-order factors	Stimulation	Affective Merit	Cognitive Merit	Stress	
Hartley & Hogan (1972)	Overall Evaluation	Student-Teacher Interaction	Structure or Organization	Load or Difficulty	
Frey (1973a)	Teacher's Presentations	Teacher Accessibility	Organization, Planning	Work Load	+ two other factors
McKeachie & Lin (1973)	Skill	(Group Interaction)	Structure	Difficulty	

[1] After J.A. Kulik and W.J. McKeachie, 1975.

it is possible to draw some tentative conclusions from the work to date.

Although several studies have found differences among ratings of male and female students (Bending, 1952; McKeachie et. al., 1971) the weight of evidence from the best designed studies shows no significant differences. Furthermore, there seems to be no complex interaction effects between student sex and instructor sex (Elsmore and Lapointe, 1974).

The student's university major seems to have no effect on ratings however university level (year) has been shown rather consistently to affect student evaluations. In most studies upper-level and graduate-level students rate their instructors and courses more favourably than lower-level students.

Perhaps the most controversial area of student evaluations is the effect of course grades or expected course grades on ratings. Unfortunately, findings in this area have been mixed. A substantial number of investigations have found significant and positive relationships between grades and evaluations (Kennedy, 1975), however an equal number of studies have reported no such effect. Costin, Greenough, and Menges (1971) summarize the research on grades and ratings as follows:

> Does the evidence, then, support an assertion that a teacher can get "good" ratings simply by assigning "good" grades, or creating the expectancy that he will do so? The fact that the positive correlations which were obtained between student ratings and grades were typically low weakens this claim as a serious argument against the validity of student ratings. The positive findings that do occur might better be viewed as a partial function of the better achieving student's greater interest and motivation, rather than as a mere contamination of the validity of student ratings.

Commenting on Costin, Greenough, and Menges conclusions, Kulik and McKeachie (1975) cite the work of Elliott (1950) and Morsh and Wilder (1954). These findings support the belief that the relationship of grades to ratings can be best viewed as the product of a complex

Table 2.8 EFFECT OF STUDENT VARIABLES ON RATINGS

Student Variable	Study	Effect
Student's Sex	Remmers (1939)	no significant effect
	Bendig (1952)	females rated less favourably
	Rayder (1968)	no significant effect
	McKeachie et. al (1971)	females rated more favourably
	Elsmore & Lapointe (1974)	no significant effect
Student's major	Cohen & Humphreys (1960)	no significant effect
	Rayder (1968)	no significant effect
University level (Year)	Remmers & Elliott (1949)	grad. students rated more favourably than undergrad. students
	Gage (1961)	students in advanced courses rated more favourably than those in lower level courses
	Miller (1972)	upper division courses were rated more favourably than lower division courses
Course Grade	Voeks & French (1960)	no significant effect
	Remmers (1960)	no significant effect
	Kennedy (1975)	students receiving 'A' or 'B' rated more favourably than those receiving 'C' or 'D'

student ability-level by teacher presentation-level interaction.

> .....if the instructor teaches for the bright students, he will be approved by them and there will be a positive correlation between ratings and grades; if he teaches for the weaker students, he will be disapproved by the bright students and a negative coefficient will be obtained. This sort of interaction could explain the diverse findings in this area reviewed by Costin, Greenough, and Menges (1971), who found that some studies report a negative correlation, some a positive correlation, and some a non-significant correlation between student ratings and grades.

2.26. Effect of Instructor Variables on Ratings

The effects of instructor rank, sex, and research productivity on student ratings have been studied rather extensively (Table 2.9). With respect to academic rank findings are somewhat mixed, however, where significant differences are found, ratings invariably favour senior faculty members (Full and or Associate professors) over junior faculty members. The instructor's sex seems to have no effect on ratings.

One topic that seems to spark considerable controversy among faculty members everywhere is the relationship between research productivity and teaching effectiveness. There are those who claim that good teaching and good research go hand-in-hand; each complementing the other. Others claim that the two activities are mutually destructive; good teachers have little time for good research and good researchers have little time for students and teaching. One position that is not often heard is that research ability and teaching ability are essentially independent traits; good teachers may or may not be good researchers and good researchers may or may not be good teachers.

The position that teaching and research are complementary activities is supported in part by the work of Bresler (1968) and McDaniel and Feldhusen (1970). Bresler found that faculty members who receive more outside funding for research purposes also receive more favourable student ratings. Unfortunately, Bresler's findings were not accompanied

Table 2.9 EFFECT OF INSTRUCTOR VARIABLES ON RATINGS

Instructor Variable	Study	Effect
Instructor Rank	Downie (1952)	Full professors rated more favourably than other ranks
	Gage (1961)	Full and Associate Professors rated more favourably than Asst. Professors and Instructors
	Langen (1966)	Decreasing favourability as follows: Assoc. Professors, Assistant Professors, Instructors (Full Prof. not studied)
	Aleamoni & Yimer (1973)	no significant correlation between rank and student ratings
Instructor Sex	Elliott (1950)	no significant effect
	Lowell & Haner (1955)	no significant effect
	Aleamoni & Yimer (1973)	no significant effect
	Elsmore & Lapointe (1974)	no significant effect
Research Productivity	Voeks (1962)	no relationship between research productivity and student ratings
	Bresler (1968)	Faculty who were more successful in receiving outside research funding received more favourable ratings
	McDaniel & Feldhusen (1970)	mixed findings (see text)
	Hayes (1971)	no relationship between research productivity and student ratings

by tests of statistical significance and varied greatly from one academic discipline to another. Bresler's research, reported in Science, was severely criticized on statistical and methodological grounds by Quereshi (1968) whose rebuttal was also published in Science.

McDaniel and Feldhusen studied the relationship of scholarly activity (as measured by: 1) number of 1st and 2nd authorships of books; 2) number of 1st and 2nd authorships of journal articles; and 3) grant status) to student ratings. Correlations were generally low, but significant and positive relationships were found between second authorship of articles and ratings. However, negative relationships were found between first authorship of articles or books and ratings. Furthermore, no differences were found in student ratings between faculty members who held a research grant and those who did not.

The work of Voeks (1962) and Hayes (1971) supports the contention that teaching effectiveness and research productivity are not related. In the Hayes study, research productivity was measured in three ways:

1) publication rate (weighted by type of publication),
2) grant status, and
3) rating by department chairman.

Teaching effectiveness was measured by:

1) average student ratings over 4 semesters, and
2) department chairman's rating of ability.

Only one of the six possible correlations between research and teaching measures was found to be significant - that between chairman's research rating and chairman's teaching rating.

In summary, it appears that if teaching effectiveness and research productivity are related, the relationship is at best a weak one. The strongest evidence seems to support the contention that the two activities are in fact unrelated.

2.27 Effect of Class Variables on Ratings

The effects of two class variables on student ratings have been studied extensively (Table 2.10). These variables are class size and course status (i.e.: whether the course is required or elective for the majority of the students enrolled).

Table 2.10 EFFECT OF CLASS VARIABLES ON RATINGS

Class Variable	Study	Effect
Class Size	Gage (1961)	curvilinear relationship; both large and small classes were rated more favourably than moderate-size classes
	McDaniel & Feldhusen (1971)	small classes were rated most favourably
	Miller (1972)	small classes were rated most favourably
	Wood et. al. (1974)	curvilinear relationship; both large and small classes were rated more favourably than moderate-size classes
	Aleamoni & Graham (1974)	no significant effect
	Crittenden et. al. (1975)	small classes were rated most favourably
Course status (compulsory/elective)	Lovell & Haner (1955)	elective courses were rated more favourably than required courses
	Cohen & Humphreys (1960)	elective courses were rated more favourably than required courses
	Gage (1961)	elective courses were rated more favourably than required courses
	Miller (1972)	no significant effect

It is widely believed that student ratings of courses and instructors are inversely related to class size. In fact, the strongest evidence tends to support this view. However, a number of studies have reported a curvilinear relationship in which small and large classes receive equally favourable evaluations and moderate-size classes receive significantly lower ratings (Gage, 1961; Wood et. al, 1974). Other studies have shown no class-size effect (Aleamoni and Graham, 1974).

In the introduction to their paper Crittenden et. al. (1975) discuss possible reasons for these inconsistent findings. Four explanations are given.

First, in many studies the sample size (number of classes) is relatively small, casting some doubt on the reliability of the results. Second, there is no agreement among studies regarding the operational definition of size categories. For example, the definition of "large" has varied from "10 or more" to "200 or more." Third, it may be that some students alter their expectations of instructional performance to take into account factors such as class size. Finally, some institutions or departments may attempt to counteract the presumed class-size effect by assigning their best instructors or allocating more resources to larger classes.

Crittenden and his associates go on to report the results of a well-designed study consisting of 981 classes at the University of Illinois at Chicago Circle. The same evaluation instrument was administered in all classes and 8 size categories were used without assigning labels to them. Class size ranged from under 20 to over 600. The findings show a clear linear relationship in which mean student ratings decrease with increasing class size.

The results of studies on the relationship of course status (compulsory/elective) to student ratings are fairly consistent. Although occasional studies report no significant effects (Miller, 1972), the weight of evidence supports the view that elective courses tend to receive more favourable ratings than required or compulsory courses. To our knowledge no study has shown that students consistently favour required courses over elective courses.

III. RELIABILITY AND VALIDITY OF THE SOST

This section presents the methods and results of the reliability and validity studies of the SOST.

3.1 INTERNAL CONSISTENCY OF THE SOST

Ss The internal consistency analyses were based on 2229 student responses without regard to class or instructor (Students were enrolled in 93 class sections taught by 53 different instructors.) Two-thirds (67.3%) of the respondents were first or second year students and the majority (64.9%) were Social Science or Science and Mathematics majors. See Table 1.3 for a further description of the students and instructors.

Analytic Methods. Cronbach's alpha coefficient was calculated for each of the subscales ("Sections") using SPSS subprogram "Reliability." Prior to the analyses several of the item scales were reversed (items 11, 16, 19, 24, 27 and 28) to insure uniform directionality.

Results. The Alpha coefficients are reported in Table 3.1. The Alphas range from .19 to .80. Internal consistencies of Sections A, B and C are moderate to relatively high and are well within the ranges reported for other teaching evaluation instruments (Table 2.2). However, the Alpha coefficients for Section D (.37) and Section E (.19) are unacceptably low.

An examination of Section D ("Feedback") by analysis of variance procedures showes that no single item is largely responsible for the low reliability. However, the deletion of item 25 ("instructor's expectations for student performance .....") would raise the Alpha coefficient to .45. Two possible explanations for this come to mind. First, the item itself seems to have little relationship to "Feedback" as do the other items, to some extent. Second, the Likert scale descriptors ("very low, low, average ....") are different from the descriptors of the remaining 3 items ("strongly agree, agree, not sure....").

As with Section D, the low internal consistency for Section E ("Standards") is not attributable to any single item. Items 26 and 27 seem to cover course workload whereas Item 28 is an evaluation of the course assignments. Deletion of Item 28 would raise the Alpha coefficient to .28.

Table 3.1 INTERNAL CONSISTENCY OF THE SOST [1,2]

Subscale	Cronbach's Alpha	Items	Alpha if Item Deleted
Section A	.78	9	.74
		10	.70
		11	.75
		12	.73
		13	.75
		14	.78
Section B	.65	15	.50
		16	.72
		17	.41
Section C	.80	18	.73
		19	.76
		20	.74
		21	.75
Section D	.37	22	.24
		23	.08
		24	.35
		25	.45
Section E	.19	26	.14
		27	-.09
		28	.28

[1] Calculation of Alphas based on 2229 student responses without regard to class or instructor.

[2] Scalings for the following items were reversed prior to data analysis: 11, 16, 19, 24, 27, and 28.

3.2 STABILITY OF THE SOST

Ss The stability analyses were based on the responses of 435 students who were enrolled in 25 sections of an introductory Psychology course (Psychology 115a). Although descriptive data for these subjects were not analyzed, the students typically represent a wide spectrum of interests, motivations and university majors.

The format of the course requires students to attend 2 weekly meetings led by a graduate teaching assistant and 1 weekly presentation by a guest lecturer. Student enrollment in sections ranged from 18 to 88 with an average enrollment of approximately 36 (35.8). The grading procedure is based on 4 objective mid-term examinations (70%) which are the same for all students enrolled in the course and a series of small projects (30%) assigned by individual section leaders.

Experimental Design and Analytic Methods

Stability by the test-retest method was examined for intervals of 7 days, 14 days, 21 days, and 28 days. The following data collection procedures were employed.

All students evaluated their section leader by completing the SOST on November 11 or 12 (depending on meeting day). A second evaluation was completed according to the following schedule:

Sections 1-6	on	November 18 or 19
Sections 7-12	on	November 25 or 26
Sections 13-18	on	December 2 and 3
Sections 19-25	on	December 9 and 10

Each student was assigned an anonymous code number. The code numbers, which were used in lieu of names or student I.D.s, permitted the matching of first and second evaluations by student. Matched pairs were obtained for 435 students.

Pearson product-moment correlation coefficients were calculated using SPSS subprogram "Pearson Corr." Within interval groups, all data was pooled and correlations were calculated without regard to section or instructor. Stabilities were examined for individual items only. The Ns associated with 7 day, 14 day, 21 day and 28 day intervals were 129, 108, 87, and 111, respectively.

Results. Stability coefficients are reported in Table 3.2. Coefficients ranged from .40 to .77 (7 day interval), .22 to .73 (14 day interval), .12 to .76 (21 day interval), and .19 to .68 (28 day interval). These coefficients are moderate to low but, with a few exceptions (asterisks), generally within the range reported for other instruments (Table 2.3).

It is generally not acceptable to compare unadjusted correlation coefficients derived from sources with varying sample sizes since the significance level of r depends on N. The calculation of mean stability coefficients across several items is also considered, by some, to be a questionable practice. It is, however, interesting to note that even with decreasing Ns the mean stability coefficients decrease as the time interval increases (compare 7 day mean to 21 day mean). This decrease is probably due to both error variance associated with the students as well as true changes in the students' perceptions of their instructors and the course.

3.3 CRITERION-RELATED VALIDITY: RELATIONSHIPS BETWEEN SOST RATINGS AND STUDENT ACHIEVEMENT

Ss These analyses were based on the responses of 620 students enrolled in 25 sections of an introductory Psychology course (Psychology 115a). The sample population was characteristically somewhat different from the total data pool (Table 3.3). The sample was composed of a relatively larger proportion of Arts and Social Science Majors (61.5%) and a smaller proportion of Science and Mathematics majors (16.9%). Over 80% of the subjects were first year students. Although the sex of the subjects was not ascertained, enrollment figures for the course generally show an equal mix of males and females.

Analytic Methods. All analyses were based on section means. The mean SOST ratings for each section were calculated using SAS procedure "Means." In addition, the mean total achievement scores for each section were calculated. The total achievement scores were expressed as percentages and represented the weighted performance on 4 multiple-choice examinations (70%) and several "subjective" projects assigned by individual section leaders (30%). Mean total achievement scores ranged from 69.4 to 79.7 for the 25 sections.

Table 3.2 STABILITY OF THE SOST[1,2]

Item	Stability Coefficients			
	7 days (N=129)	14 days (N=108)	21 days (N=87)	28 days (N=111)
9	.50	.63	.60	.50
10	.62	.65	.56	.61
11	.40	.51	.46	.33
12	.71	.67	.76	.39
13	.50	.56	.36	.40
14	.77	.39	.67	.62
15	.65	.59	.24**	.36
16	.49	.44	.48	.23***
17	.63	.63	.42	.27***
18	.73	.73	.61	.50
19	.63	.54	.46	.35
20	.63	.62	.46	.55
21	.49	.63	.58	.59
22	.43	.48	.12*	.19**
23	.65	.50	.55	.26***
24	.40	.22***	.46	.44
25	.46	.56	.49	.47
26	.63	.56	.62	.68
27	.65	.31	.54	.47
28	.54	.43	.28***	.35
Mean (all items)	.58	.53	.49	.43

[1]Because of differences in N among groups Γs should not be compared across columns.

[2]All Γs are significant at $p < .001$ with exception of asterisks (***$p < .01$ **$p < .05$, *$p > .05$).

Table 3.3 PROFILE OF INTRODUCTORY PSYCHOLOGY STUDENTS
(N=620)

Item	Frequencies				
	A	B	C	D	E
1. My major is in:	Arts	Soc. Sci.	Sci. & Math	Business	Other
	21%	40.5%	16.9%	6.7%	15.0%
2. This course is part of my:	Hon. Prgm	Gen. Prgm			
	41.9%	58.1%			
3. I have completed the following number of University level full courses:	0-2	3-7	8-12	13-17	18--
	82.4%	13.1%	2.8%	1.6%	0.2%
4. Rating myself against the performance of other students in the class, I see myself in one of the following groups.	Superior	Above Avg.	Average	Below Avg.	Failing
	3.9%	36.5%	51.9%	7.1%	0.6%
5. This course was compulsory.	Yes	No	Not Sure		
	40.7%	54.5%	4.9%		
6. My attendance and punctuality have been consistently good.	Yes	No			
	90.0%	10.0%			
7. Compared to other courses I have taken, I consider my effort in this course to have been:	Excellent	Above Avg.	Average	Below Avg.	Poor
	8.5%	40.3%	42.3%	7.6%	1.3%
8. I have found the material in this course to be inherently difficult.	Yes	No			
	24.2%	75.8%			

Pearson product-moment correlations were calculated between mean SOST ratings and mean total achievement scores across the 25 sections using SPSS subprogram "Pearson Corr."

Results. The correlation coefficients are reported in Table 3.4. Eleven (11) of the 20 coefficients are significant at the .05 level or better. This number of significant correlations is considerably greater than would be expected by chance alone.

Of the 11 significant correlations, 10 are found among items in Sections A, B and C of the SOST. This would argue that the instructor's ability to communicate with and motivate students is more important in promoting learning than the assignments, workload or evaluation system employed in the course. It further argues that good teachers (those who promote learning in their students) receive good evaluations and that poor teachers receive poor evaluations.

An examination of individual coefficients shows that, although many are statistically significant, the absolute values are moderate to low. These findings are consistent with much previous work on the subject (Table 2.6).

It will be noted that many of the significant correlations in Table 3.4 are negative. However, an inspection of the item scales shows that, where negative correlations are indicated, a low scale score (A or B) implies agreement with a generally positive statement. Furthermore, where significant correlations are positive, a high scale score (D or E) implies disagreement with a generally negative statement. In sum, regardless of the direction of the correlation coefficient (+ or -), all significant coefficients imply a positive relationship between teaching effectiveness or course structure and student achievement.

The largest correlation coefficients are associated with Items 21 ("The instructor was successful in making difficult material understandable."), 18 ("The instructor made this course as interesting as the subject matter would allow."), 10 ("The instructor presented material in a coherent manner, emphasizing major points and making relationships clear.") and 9 ("The instructor is clear and audible."). All of these items seem to be related to the instructor's general ability to communicate.

Table 3.4 RELATIONSHIPS BETWEEN SOST RATINGS AND STUDENT ACHIEVEMENT [1,2,3]

Item	Correlation Between Item and Total Achievement Score (r)
9	-.55**
10	-.56**
11	.37*
12	-.29
13	-.42*
14	.02
15	-.43*
16	.02
17	-.42*
18	-.57***
19	.38*
20	-.43*
21	-.58***
22	-.31
23	.07
24	.36*
25	.10
26	.15
27	.10
28	-.09

1 ***$p < .001$, **$p < .01$, *$p < .05$

2 N.B.: In some cases a negative correlation implies a positive relationship because of the direction of the item scale (see Results section 3.3).

3 Item responses were coded as follows: A=1, B=2, C=3, D=4 and E=5

3.4 FACTOR ANALYSIS OF THE SOST

Ss The factor analysis was based on the responses of 2229 students. For a description of these subjects see Table 1.3.

Analytic Methods. The factor analysis was performed by SPSS subprogram "Factor" with the PA2 factoring method using varimax rotation. This procedure calculates a principal-component solution with iteration and employs orthogonal rotation with Kaiser Normalization. The factoring method replaces the main diagonal elements of the correlation matrix with communality estimates and employs an iteration procedure for improving the estimates of communality.

The eigenvalue criterion for establishing the number of components (factors) was 1. To simplify interpretation and minimize the number of cross-loadings, only loadings of .40 or greater were interpreted.

Results. Five factors emerged from the analysis (Table 3.5). These factors accounted for 55.3% of the variance in the data. The first factor by itself accounted for 30.0% of the total variance.

The communalities (total variance of an item accounted for by the combination of all common factors) ranged from .08 (item 27) to .68 (item 10). The average was approximately .40 (.402).

The factorial complexity of the rotated matrix was relatively high. A number of items loaded significantly on at least two factors. An interpretation of the factor structure is complicated by these cross-loadings (Table 3.6).

Factor I: Instructional Skill

The first factor is a measure of the instructor's general ability to communicate with and motivate students. Items with the highest loadings (10, 18, 21 and 9) assess the instructor's coherence and clarity of presentation and his success in making the subject matter interesting and understandable.

Factor II: Interaction

The second factor seems to relate to student-teacher rapport and the general level of verbal and written exchanges between the instructor and the student.

Table 3.5 FACTOR ANALYSIS OF SOST
(N=2229)

Item	Communality	Varimax Rotated Factor Matrix				
		Factor I	Factor II	Factor III	Factor IV	Factor V
9	.44	.60	.18	.00	.21	-.09
10	.68	.77	.18	.03	.22	-.09
11	.38	-.52	-.12	-.05	-.19	.25
12	.45	.48	.07	.06	.46	-.05
13	.38	.35	.24	.07	.40	-.21
14	.30	.17	.12	-.03	.50	.06
15	.41	.32	.39	-.12	.37	-.01
16	.20	-.12	-.31	.13	-.25	.11
17	.48	.29	.50	-.07	.38	.03
18	.57	.63	.37	-.03	.16	-.03
19	.48	-.47	-.37	.26	-.04	.21
20	.51	.43	.55	-.12	.10	.05
21	.57	.61	.42	.03	.11	-.04
22	.39	.10	.60	.12	.11	-.01
23	.30	.15	.38	.23	.13	-.24
24	.51	-.14	-.15	.13	-.12	.66
25	.24	-.05	.01	.49	-.01	.01
26	.31	.06	.04	.55	-.03	-.06
27	.08	-.04	.05	-.11	.07	.25
28	.36	.21	.56	.06	.05	-.03

Table 3.6 FACTOR STRUCTURE OF SOST.

Factor	Factor Loadings	Items
I Instructional skill	.60	9. The Instructor is clear and audible.
	.77	10. The Instructor presented material in coherent manner .....
	.52	11. Course material was disorganized and hindered understanding.
	.48	12. The Instructor was consistently prepared for class.
	.53	18. The Instructor made this course as interesting as the subject matter would allow.
	.47	19. The Instructor did not increase my interest in the subject matter .....
	.43	20. The Instructor motivated me to put forth a good effort.
	.61	21. The Instructor was successful in making difficult material understandable.
II Interaction	.50	17. The Instructor maintained a generally helpful attitude toward students .....
	.55	20. The Instructor motivated me to put forth a good effort.
	.42	21. The Instructor was successful in making difficult material understandable.
	.60	22. Verbal or written comments on assignments have been constructive.
	.56	28. The assignments provided a valuable learning experience.
III Workload	.49	25. The Instructor's expectations for student performance were .....
	.55	26. The amount of work required for this course has been .....
IV Organization	.46	12. The Instructor was consistently prepared for class.
	.40	13. The Instructor was clear on what was expected .....
	.50	14. The Instructors attendance and punctuality have been consistently good.
V Feedback	.66	24. Throughout this course, I have not been able to assess my progress and achievement.

Factor III: Workload

The third factor is a measure of the amount of student effort required by the instructor and the course. The two items with high loadings on this factor (25 and 26) assess the amount of work required for the course and the instructor's expectations for student performance.

Factor IV: Organization

The fourth factor is an assessment of the organizational skills of the instructor. Items loading highly on this factor (12, 13, and 14) pertain to the instructor's attendance and punctuality, his preparedness, and the clarity with which he has stated his objectives and requirements.

Factor V: Feedback

The fifth factor (Item 24) is a measure of the extent to which the student is able to judge his level of performance in the course.

3.5 THE EFFECT OF STUDENT VARIABLES ON SOST RATINGS

Ss These analyses were based on the entire data pool consisting of 2229 student responses regardless of class section or instructor (except where noted). For a description of these subjects see Table 1.3.

Analytic Methods. Analysis of variance procedures were used to examine the effects of the following variables on student ratings: (1) the student's major (faculty affiliation - Item 1); (2) the student's level (number of courses completed - Item 3); (3) the student's perception of his own performance relative to other students in the class (Item 4); (4) whether the course was compulsory or elective (Item 5); (5) the student's perception of his effort in the course relative to other courses he has taken (Item 7).

A series of one-way multivariate and univariate analyses of variance were performed using the "General Linear Models" (GLM) procedure of the SAS package. A fixed-effects model (I) was used. Only the univariate F-ratios and their associated significance levels are reported. Post hoc analyses were not performed, however, means and standard deviations were calculated on anova levels using SPSS subprogram "Breakdown."

3.51 The Effect of Student's Major on SOST Ratings

All analyses of variance showed significant differences in SOST

ratings by student major (Table 3.7). Business majors rated their instructors and courses most favourably on 12 of the 20 items. Arts majors rated most favourably on 6 of the items and Arts/Business majors rated identically on 2 items.

Least favourable ratings were given on 12 items by students who identified their major as "other." Science and Mathematics majors rated their instructors and courses least favourably on 6 items.

Although no simple and totally consistent pattern emerged from these analyses, it appears that Business majors are more lenient (favourable) in their evaluations than non-Business students. Furthermore, Science and Mathematics students and "others" tend to be harshest (least favourable) in their ratings.

It should be pointed out that Business students constituted the smallest group in the sample population (8.8%) and that approximately 30% of these ratings were obtained in only 2 class sections. In addition, the Science and Mathematics students comprised that largest group (40.4%) and evaluated the largest number of courses and instructors. Although no tests for homogeneity of variances were performed, heterogeneity might account for some of the observed differences. This explanation is unlikely, however, since F is known to be robust with respect to departures from homoscedasticity (Winer, 1971).

3.52 The Effect of Student's Level on SOST Ratings

Thirteen (13) of the 20 SOST items showed significant differences by student level (Table 3.8). As with student major, no clear and consistent pattern is discernable on the basis of the number of courses completed by the student, although a few generalizations can be made.

Upper-level students (those having completed at least 13 courses) tended to rate their instructors more favourably than lower-level students in terms of ability to communicate and motivate (Items 9, 18, and 20). In addition, upper-level students were more inclined to rate the instructor's expectations (Item 25) as high and the course workload (Item 26) as relatively heavy. Finally, honours and graduate-level students (18 or more courses) considered their instructor's punctuality to have been better than other students (Item 14) however, they considered

(N=2229)

Item	Means and (Standard Deviations) by Major					F ratio[2]
	Arts (A)	Social Science (B)	Science and Math (C)	Business (D)	Other (E)	
9	1.76 (.80)	1.84 (.86)	1.93 (.90)	1.63 (.73)	2.24 (1.13)	17.08
10	2.10 (1.05)	2.19 (1.08)	2.34 (1.06)	1.89 (.83)	2.43 (1.09)	12.41
11	3.92 (.98)	3.82 (1.02)	3.65 (1.03)	4.00 (.89)	3.66 (.99)	7.20
12	1.79 (.88)	1.83 (.83)	1.97 (.90)	1.62 (.75)	1.87 (.76)	9.33
13	1.76 (.97)	1.86 (.96)	2.08 (.99)	1.74 (.74)	2.02 (.97)	11.73
14	1.35 (.59)	1.41 (.68)	1.53 (.67)	1.28 (.54)	1.56 (.74)	11.37
15	1.66 (.77)	1.75 (.86)	1.93 (.94)	1.72 (.80)	2.07 (.97)	12.60
16	3.76 (.93)	3.68 (.89)	3.60 (.96)	3.57 (.98)	3.51 (.89)	3.70**
17	1.86 (.84)	1.97 (.88)	2.02 (.87)	1.94 (.81)	2.21 (.90)	6.53
18	2.23 (1.11)	2.32 (1.10)	2.37 (1.02)	2.07 (.97)	2.66 (1.16)	10.19
19	3.43 (1.15)	3.22 (1.15)	3.24 (1.12)	3.34 (1.10)	2.98 (1.20)	6.09
20	2.65 (1.00)	2.78 (1.03)	2.76 (1.00)	2.54 (.99)	3.10 (1.04)	10.53
21	2.31 (1.00)	2.38 (.93)	2.53 (.95)	2.31 (.86)	2.66 (1.13)	6.93
22	2.49 (.94)	2.58 (.87)	2.70 (.94)	2.69 (.91)	2.79 (.88)	4.31**
23	2.22 (.98)	2.38 (1.02)	2.60 (1.04)	2.37 (.89)	2.43 (.99)	7.40
24	3.86 (.95)	3.82 (.93)	3.58 (.98)	3.45 (1.04)	3.52 (1.02)	11.17
25	3.48 (.75)	3.48 (.67)	3.56 (.75)	3.58 (.85)	3.40 (.87)	2.73*
26	3.42 (1.05)	3.48 (1.02)	3.50 (.86)	3.68 (1.02)	3.63 (1.03)	3.23*
27	2.78 (1.29)	2.77 (1.21)	2.93 (1.21)	2.51 (1.23)	2.76 (1.22)	4.52**
28	2.26 (.95)	2.43 (.99)	2.47 (.96)	2.26 (.86)	2.65 (1.05)	5.94

[1] The number of missing cases varied by item with a range of 0.5% to 4.7%; Based on 2217 responses (0.5% missing) the breakdown of responses by major was: Arts (13.1%); Social Science (24.6%); Science and Math (40.4%); Business (8.8%); Other (13.1%)

[2] All Fs are significant at $p < .0001$ with exception of asterisks (**$p < .01$, *$p < .05$)

Table 3.8 THE EFFECT OF STUDENT'S LEVEL ON SOST RATINGS[1]

(N=2229)

Item	Means and (Standard Deviations) by Level					F ratio[2]
	0-2 (A)	3-7 (B)	8-12 (C)	13-17 (D)	18-- (E)	
9	1.88 (.84)	1.99 (.96)	2.13 (1.06)	1.84 (1.02)	1.72 (.90)	8.02
10	2.26 (1.04)	2.22 (1.05)	2.35 (1.06)	2.13 (1.14)	2.22 (1.09)	1.24(NS)
11	3.80 (.95)	3.70 (1.05)	3.68 (1.04)	3.77 (1.11)	3.69 (1.10)	1.90(NS)
12	1.91 (.85)	1.83 (.80)	1.81 (.80)	1.65 (.81)	1.91 (.96)	4.72***
13	1.90 (.93)	1.93 (.91)	1.99 (.93)	1.80 (.90)	2.20 (1.17)	6.86
14	1.46 (.67)	1.47 (.67)	1.53 (.74)	1.48 (.64)	1.36 (.57)	2.52*
15	1.85 (.91)	1.82 (.88)	1.98 (.96)	1.79 (.87)	1.80 (.87)	1.90(NS)
16	3.57 (.87)	3.66 (.92)	3.73 (.96)	3.95 (.96)	3.55 (1.11)	5.94
17	2.03 (.85)	2.00 (.88)	2.09 (.91)	1.90 (.88)	1.94 (.89)	2.05(NS)
18	2.40 (1.06)	2.41 (1.09)	2.41 (1.13)	2.09 (1.05)	2.22 (1.07)	5.97
19	3.20 (1.12)	3.18 (1.19)	3.26 (1.18)	3.46 (1.18)	3.29 (1.15)	1.79(NS)
20	2.88 (.99)	2.80 (1.07)	2.73 (1.01)	2.44 (1.06)	2.61 (1.02)	11.44
21	2.47 (.96)	2.46 (.93)	2.55 (1.02)	2.32 (1.05)	2.44 (1.02)	1.43(NS)
22	2.69 (.86)	2.64 (.93)	2.59 (.94)	2.51 (1.00)	2.65 (1.03)	3.24*
23	2.47 (1.02)	2.40 (.99)	2.40 (.97)	2.26 (1.00)	2.64 (1.06)	3.80**
24	3.72 (.96)	3.63 (1.02)	3.58 (.96)	3.63 (1.00)	3.50 (1.05)	5.70
25	3.42 (.66)	3.47 (.77)	3.38 (.82)	3.70 (.79)	3.88 (.83)	17.22
26	3.54 (.95)	3.48 (.98)	3.26 (.98)	3.55 (.96)	3.66 (.97)	5.77
27	2.69 (1.22)	2.93 (1.26)	3.02 (1.18)	2.95 (1.22)	2.86 (1.23)	4.92***
28	2.45 (.93)	2.45 (1.01)	2.50 (1.02)	2.38 (1.15)	2.40 (.97)	0.76(NS)

The number of missing cases ranged from 1.2% to 5.2%; Based on 2202 responses (1.2% missing) the breakdown of responses by levels was: 0-2 (52.0%); 3-7 (15.3%); 8-12 (10.5%); 13-17 (8.4%); 18- (13.7%).

All Fs are significant at $p < .001$ with exceptions of (NS) and asterisks (***$p < .001$, **$p < .01$, *$p < .05$)

the evaluation system less fairly applied (Item 23), their ability to assess their own progress and achievement less marked (Item 24), and the instructors expectations less clearly delineated (Item 13).

In general these findings are supportive of previous work which has shown that upper-level students tend to evaluate instructors' presentations more positively (Table 2.8).

3.53 The Effect of Student's Performance on SOST Rating

Sixteen (16) of the 20 SOST items showed significant differences when responses were classified by the student's perception of his own performance relative to other students in the class (Table 3.9). These findings are interesting on several accounts.

Quite aside from the question at hand, the percentages of students who classify themselves in each category is at least of passing interest. Over 88% of the students see themselves as "average" or "above average." Almost 5% classify their performance as "superior" and only 1% see themselves failing. It might be interesting to compare these self-appraisals with grades actually received in courses. It would appear that students tend to cluster themselves in the centre of a grade distribution, perhaps to a greater extent than their professors do. Our guess is that it is a rare professor (in these days of "grade inflation") who assigns only 5 "As" and 1 "F" in a class of 100 students.

The results of the analyses of variance are equally interesting. A clear and fairly consistent pattern indicates that students who see themselves as "below average" or "failing" tend to rate their instructor and the course less favourably than other students. Equally consistent findings show that students who perceive their performance as "above average" or "superior" rate their instructors as more effective, the feedback as "constructive," the evaluation system as "fairly applied," and the assignments as a "valuable learning experience." It appears then that a direct (and perhaps linear) relationship exists between students' perceptions of their own performance relative to others in the class and their evaluation of the instructor and the course.

These findings are probably not surprising to many, however, it is interesting that similar findings have not been widely reported (to our knowledge). The findings, furthermore, tend to cast some doubt on

Table 3.9 EFFECT OF STUDENT'S PERFORMANCE ON SOST RATINGS[1]
(N=2229)

Item	Means and (Standard Deviations) by Performance					F ratio[2]
	Superior (A)	Above Average (B)	Average (C)	Below Average (D)	Failing (E)	
9	1.89 (1.14)	1.82 (.89)	1.94 (.90)	2.06 (.94)	2.05 (.95)	3.53**
10	2.13 (1.12)	2.18 (1.07)	2.28 (1.03)	2.40 (1.18)	2.55 (1.18)	2.91*
11	3.84 (1.14)	3.83 (1.04)	3.73 (.96)	3.57 (1.04)	3.24 (1.26)	4.53**
12	1.75 (.87)	1.84 (.89)	1.90 (.83)	1.89 (.84)	1.77 (.61)	1.11(NS
13	1.83 (.92)	1.89 (.97)	1.99 (.95)	2.05 (1.07)	2.23 (1.02)	2.47*
14	1.33 (.60)	1.45 (.69)	1.49 (.67)	1.43 (.54)	1.27 (.55)	2.30(NS
15	1.81 (.93)	1.80 (.87)	1.88 (.90)	1.82 (.91)	2.32 (1.21)	2.61*
16	3.73 (.96)	3.66 (.96)	3.60 (.92)	3.52 (.93)	3.55 (.80)	1.45(NS
17	2.01 (.94)	1.93 (.83)	2.05 (.88)	2.11 (.98)	2.50 (.86)	5.59***
18	2.28 (1.14)	2.27 (1.06)	2.41 (1.07)	2.42 (1.08)	2.64 (1.26)	3.16*
19	3.44 (1.24)	3.35 (1.11)	3.17 (1.15)	2.98 (1.23)	2.64 (1.00)	5.73***
20	2.62 (1.05)	2.67 (1.02)	2.82 (1.00)	3.04 (1.01)	3.33 (1.15)	6.47***
21	2.27 (1.00)	2.37 (.98)	2.50 (.94)	2.81 (1.04)	2.95 (1.13)	10.04***
22	2.61 (.93)	2.58 (.90)	2.69 (.91)	2.73 (.99)	3.18 (1.05)	4.66***
23	2.30 (.95)	2.38 (.95)	2.47 (1.00)	2.77 (1.09)	3.23 (1.27)	9.01***
24	3.91 (1.04)	3.85 (.91)	3.55 (.98)	3.16 (1.11)	2.95 (1.25)	20.48***
25	3.49 (.84)	3.53 (.78)	3.52 (.74)	3.44 (.67)	3.23 (1.11)	1.22(NS
26	3.30 (1.03)	3.44 (.97)	3.57 (.94)	3.76 (.97)	3.59 (1.33)	6.71***
27	2.94 (1.28)	2.94 (1.25)	2.75 (1.19)	2.41 (1.18)	2.55 (1.50)	5.54***
28	2.24 (.86)	2.42 (.96)	2.45 (.97)	2.56 (1.11)	2.90 (1.14)	2.40*

The number of missing cases ranged from 0.7% to 4.8%; Based on 2213 responses (0.7% missing) the breakdown of responses by performance was: Superior (4.9%), Above Avg. 38.5%), Average (49.6%), Below Avg. (6.1%), Failing (1.0%).

****$p < .0001$, ***$p < .001$, **$p < .01$, *$p < .05$

the validity of individual student ratings though not necessarily on ratings received by an instructor from an entire class.

3.54 The Effect of Course Status (Compulsory/Elective) on SOST Rating

Prior to analysis the responses of students who were "not sure" of their course status were dropped from the data pool resulting in 2139 useable responses (95.9% of original data pool). Out of convenience, the analysis of variance procedure was used even though the "Student's t-test" is more often employed in a 2-group design. t is nothing more than a "step-down" of F and both yield findings having identical "significance levels."

Significant differences were found for 18 of the 20 SOST items when responses were classified by course status (Table 3.10). In every case, students rated "elective" courses more favourably than "compulsory" courses. The two items for which no differences were found asked students to assess the instructor's attendance and puntuality (item 14) and his expectations for student performance (Item 25).

The finding that "elective" courses are more attractive to students than "compulsory" courses is not particularly surprising. It does, however, again question the validity of individual student ratings and, in some cases, even class ratings. These results are generally consistent with previous research (Table 3.10).

3.55 The Effect of Student's Effort on SOST Ratings

Significant differences were found among 16 of the 20 SOST items when responses were classified by the student's perception of his own effort relative to his effort in other courses (Table 3.11). Where differences were found, a consistent patten of ratings by effort emerged.

Students who reported their effort as "excellent" or "above average" consistently rated the instructor and the course more favourably than other students. Moreover, those who indicated that their effort was "below average" or "poor" gave the least favourable evaluations.

One might profitably speculate about the relationship of student effort to teaching evaluations. It might be that students who "try harder" are more likely to succeed and thereby see the instructor and

Table 3.10 EFFECT OF COURSE STATUS (COMPULSORY/ELECTIVE ON SOST RATINGS (N=2139)[1]

Item	Means and (Standard Deviations) by Course Status		F ratio[2]
	Compulsory (A)	Elective (B)	
9	2.00 (1.00)	1.77 (.78)	33.73
10	2.33 (1.09)	2.16 (1.01)	13.82**
11	3.64 (1.05)	3.89 (.96)	31.37
12	1.92 (.86)	1.81 (.85)	7.99**
13	2.06 (1.00)	1.83 (.91)	31.85*
14	1.47 (.65)	1.44 (.67)	0.71(NS)
15	1.95 (.97)	1.73 (.79)	24.79
16	3.55 (.95)	3.72 (.90)	14.65
17	2.09 (.92)	1.92 (.79)	19.75
18	2.44 (1.10)	2.24 (1.04)	16.20
19	3.09 (1.15)	3.39 (1.12)	40.23
20	2.84 (1.04)	2.71 (.99)	6.77**
21	2.57 (1.00)	2.35 (.93)	22.85
22	2.77 (.93)	2.53 (.89)	25.38
23	2.62 (1.03)	2.27 (.97)	51.14
24	3.54 (1.03)	3.79 (.92)	32.98
25	3.53 (.79)	3.50 (.72)	0.00(NS)
26	3.66 (.92)	3.37 (1.00)	39.60
27	2.75 (1.22)	2.86 (1.24)	4.00*
28	2.53 (.99)	2.34 (.94)	17.58

[redacted]sponses of students who were "not sure" of course status were dropped from this analysis. Of the remaining 2139 respondents 53.6% evaluated a "compulsory" course, and 45.8% evaluated an "elective". Missing cases constituted the remaining 0.6%.

[2] All Fs significant at $p<.0001$ except asterisks (***$p<.001$, **$p<.01$, *$p<.05$,

Table 3.11 THE EFFECT OF STUDENT'S EFFORT ON SOST RATINGS[1]
(N=2229)

Item	Means and (Standard Deviations) by Effort					F ratio[2]
	Excellent (A)	Above Average (B)	Average (C)	Below Average (D)	Poor (E)	
9	1.79 (.98)	1.83 (.89)	1.96 (.89)	2.03 (.96)	2.35 (1.23)	4.78***
10	2.12 (1.18)	2.20 (1.06)	2.26 (1.00)	2.48 (1.10)	2.92 (1.35)	5.46***
11	3.63 (1.26)	3.79 (1.03)	3.77 (.92)	3.67 (.99)	3.81 (.98)	1.53(NS)
12	1.67 (.87)	1.83 (.87)	1.91 (.80)	1.97 (.92)	2.42 (1.06)	6.99
13	1.93 (1.07)	1.92 (.98)	1.98 (.94)	1.94 (.90)	2.19 (.98)	0.77(NS)
14	1.31 (.61)	1.41 (.62)	1.52 (.69)	1.55 (.78)	1.73 (.92)	7.40
15	1.71 (.95)	1.83 (.90)	1.89 (.89)	1.90 (.84)	2.12 (1.07)	2.27(NS)
16	3.70 (1.07)	3.68 (.94)	3.58 (.89)	3.53 (.89)	3.42 (1.10)	3.40**
17	1.82 (1.00)	1.96 (.86)	2.06 (.83)	2.14 (.86)	2.62 (1.17)	7.11
18	2.13 (1.15)	2.27 (1.05)	2.44 (1.05)	2.56 (1.14)	2.73 (1.31)	6.22
19	3.42 (1.23)	3.40 (1.12)	3.11 (1.12)	2.89 (1.17)	2.54 (1.07)	14.82
20	2.26 (1.10)	2.62 (1.01)	2.95 (.94)	3.29 (.84)	3.56 (1.00)	40.63
21	2.20 (1.00)	2.39 (.96)	2.54 (.96)	2.77 (.96)	2.81 (1.13)	11.34
22	2.43 (1.00)	2.61 (.89)	2.70 (.90)	2.85 (.93)	3.17 (1.00)	6.71
23	2.40 (1.15)	2.41 (1.03)	2.49 (.98)	2.46 (.90)	2.84 (1.21)	1.38(NS)
24	3.86 (1.02)	3.75 (.97)	3.58 (.95)	3.39 (1.07)	3.12 (1.21)	9.08
25	3.86 (.94)	3.57 (.75)	3.42 (.69)	3.29 (.76)	3.31 (.84)	13.05
26	3.99 (1.04)	3.74 (.88)	3.31 (.86)	2.94 (1.16)	2.65 (1.23)	55.02
27	2.56 (1.31)	2.73 (1.22)	2.91 (1.19)	2.94 (1.22)	3.35 (1.41)	6.25
28	2.12 (1.01)	2.35 (.93)	2.52 (.96)	2.84 (1.05)	3.00 (1.12)	14.98

The number of missing cases ranged from 0.7% to 4.8%; Based on 2214 responses (0.7% missing) the breakdown of responses by effort was: Excellent (10.3%), Above Avg. (39.9%), Average (41.6%), Below Avg. (7.0%), Poor (1.2%).

All Fs are significant at $p<.0001$ with exception of asterisks (***$p<.001$, **$p<.01$)

his course in a more favourable light. This, however, is pure speculation and the findings are insufficient to support such a causal relationship.

To our knowledge, findings of this sort have not been widely reported on the literature. The results again question the notion that individual student ratings are not biased by presumably irrelevant factors.

3.6 THE EFFECT OF INSTRUCTOR VARIABLES ON SOST RATINGS

Ss These analyses were based on the responses of 2229 students who were enrolled in 93 class sections taught by 53 different instructors. For a description of these subjects see Table 1.3.

Analytic Methods. Analysis of variance procedures were used to assess the effects of the instructor's rank and sex on SOST ratings. All analyses were performed on class means (N=93) since mean ratings are most often used to assess teaching effectiveness.

Mean ratings for individual items within class sections were calculated using SAS procedure "means." One-way analyses of variance were performed on class ratings by instructor's rank and by instructor's sex. The "GLM" procedure of the SAS package was used in these analyses. A fixed-effects model (I) was employed. Only univariate F-ratios and their significance levels are reported. Student "t-tests" were not performed even though they are commonly used in 2-group designs. See section 3.54. A breakdown of means and standard deviations by sex and rank was accomplished using SPSS subprogram "Breakdown."

A description of instructors by sex and rank is presented in Table 1.2.

3.61 The Effect of Instructor's Rank on SOST Ratings

For purposes of this analysis 4 academic ranks were identified: Professor, Associate Professor, Assistant Professor, and Other. The category of "other" includes all non-professorial teaching staff including Lecturers, Instructors, and Teaching Assistants.

Five (5) of the SOST items showed significant differences when class means were categorized by the academic rank of the instructor. In all cases the evaluations tended to favour senior staff members.

(Professor, Associate Professor) over junior staff members (Table 3.12).

Senior staff members were judged to be more consistently prepared for class (Item 12), more readily available for consultation (Item 16), generally more helpful in their attitude toward students (Item 17), better able to motivate students (Item 20), and less demanding in the amount of work required (Item 26). These findings may be surprising to some, however, they are consistent with previous work in the area (Table 2.9).

It had been our belief and that of many others that students perceive senior faculty as far too busy with research, and professional matters to be available and helpful to undergraduate students. Interestingly, this turns out not to be the case.

These findings, however, do not address a more important question: "Do favourable ratings imply that senior faculty members are in fact more effective teachers than junior faculty members"? Or, put another way, "Are students biased in their ratings with respect to their instructors' age, experience, demeanor and general appearance or do they in fact learn more effectively when taught by senior faculty members"? Tantalizing as this question is, it is simply unanswerable on the basis of the available evidence.

3.62 The Effect of Instructor's Sex on SOST Ratings

Significant differences were found on 7 SOST items when class section means were classified by instructor's sex. Where differences were found, mean ratings consistently favoured male instructors over female instructors (Table 3.13). These findings are not consistent with a large body of evidence which tends to show that student ratings are not affected by instructor's sex.

Logically, one might entertain 3 possible explanations for these findings:

1) the male instructors in the sample were in fact more effective teachers than the female instructors
2) the students were biased in their evaluations
3) the effects of instructor sex are confounded by other variables.

Table 3.12 EFFECT OF INSTRUCTOR'S RANK ON SOST RATINGS[1]
(N=93)

Item	Means and (Standard Deviations) by Rank				F ratio[2]
	Professor	Associate Professor	Assistant Professor	Other	
9	1.99 (.76)	1.78 (.57)	1.74 (.39)	1.84 (.32)	0.81(NS)
10	2.04 (.66)	2.21 (.61)	2.11 (.71)	2.23 (.45)	0.48(NS)
11	3.95 (.54)	3.78 (.72)	3.66 (.64)	3.81 (.40)	0.70(NS)
12	1.64 (.53)	1.77 (.34)	1.86 (.73)	2.06 (.45)	2.70*
13	1.73 (.60)	2.06 (.51)	1.98 (.45)	1.98 (.44)	1.06(NS)
14	1.24 (.19)	1.54 (.31)	1.47 (.33)	1.54 (.41)	1.95(NS)
15	1.64 (.36)	1.69 (.51)	1.89 (.45)	1.71 (.32)	1.42(NS)
16	3.98 (.48)	3.97 (.51)	3.62 (.48)	3.57 (.24)	6.42***
17	1.64 (.38)	1.77 (.47)	2.10 (.40)	1.84 (.27)	4.87**
18	2.05 (.60)	2.08 (.59)	2.26 (.56)	2.37 (.43)	1.95(NS)
19	3.55 (.54)	3.46 (.48)	3.31 (.52)	3.24 (.44)	1.70(NS)
20	2.34 (.51)	2.50 (.56)	2.60 (.47)	2.81 (.38)	4.31**
21	2.23 (.56)	2.35 (.52)	2.37 (.54)	2.33 (.36)	0.21(NS)
22	2.13 (.50)	2.49 (.44)	2.44 (.49)	2.46 (.40)	1.70(NS)
23	2.15 (.58)	2.29 (.35)	2.41 (.58)	2.55 (.44)	2.59(NS)
24	3.55 (.52)	3.60 (.38)	3.45 (.45)	3.70 (.34)	2.03(NS)
25	3.51 (.24)	3.53 (.25)	3.49 (.50)	3.42 (.30)	0.55(NS)
26	2.81 (.48)	3.35 (.50)	3.61 (.51)	3.71 (.36)	13.32****
27	2.99 (.55)	2.90 (.24)	2.88 (.54)	2.74 (.43)	1.37(NS)
28	2.13 (.54)	2.20 (.35)	2.29 (.36)	2.33 (.28)	1.20(NS)

[1] Of the 93 class sections, 10 were taught be Professors, 14 were taught by Associate Professor, 20 were taught by Assistant Professors, and 49 were taught by "other" staff members, primarily T.A.s.

[2] ****$p<.0001$, ***$p<.001$, **$p<.01$, *$p<.05$.

Table 3.13 EFFECT OF INSTRUCTOR'S SEX ON SOST RATINGS[1]
(N=93)

Item	Means and (Standard Deviations) by Sex		F ratio[2]
	Male	Female	
9	1.76 (.34)	1.92 (.52)	3.35(NS)
10	2.09 (.56)	2.30 (.54)	3.26(NS)
11	3.86 (.50)	3.69 (.55)	2.16(NS)
12	1.87 (.62)	2.01 (.37)	1.46(NS)
13	1.82 (.36)	2.17 (.54)	13.28***
14	1.45 (.35)	1.55 (.40)	1.56(NS)
15	1.69 (.39)	1.81 (.38)	2.27(NS)
16	3.75 (.46)	3.59 (.32)	3.84(NS)
17	1.85 (.41)	1.89 (.30)	0.22(NS)
18	2.16 (.51)	2.42 (.48)	6.47*
19	3.42 (.47)	3.18 (.47)	5.74*
20	2.57 (.46)	2.81 (.44)	6.22*
21	2.29 (.44)	2.39 (.46)	1.17(NS)
22	2.36 (.48)	2.52 (.36)	3.21(NS)
23	2.28 (.46)	2.67 (.45)	16.17****
24	3.65 (.39)	3.57 (.40)	0.98(NS)
25	3.45 (.35)	3.48 (.32)	0.11(NS)
26	3.44 (.51)	3.67 (.47)	4.99*
27	2.83 (.46)	2.81 (.43)	0.04(NS)
28	2.21 (.32)	2.38 (.35)	5.78*

[1] Of the 93 class sections 55 were taught by male instructors and 38 were taught by female instructors.

[2] ****$p < .0001$, ***$p < .001$, *$p < .05$.

The third explanation is perhaps the most likely answer.

An examination of Table 1.2 shows that the large majority of the female instructors in the sample are found in the lower academic ranks (as they are at the University, in general). As a result, the effect of instructor sex may be confounded by the effect of academic rank. In order to test this hypothesis one could perform a 2-way analysis of variance thereby partialling out the variance attributable to each of the main effects (sex and rank). Unfortunately, the sample size is too small for an adequate analysis. Such an analysis would be based on an experimental design containing a number of near-empty cells.

3.7 THE EFFECT OF CLASS VARIABLES ON SOST RATINGS

Ss These analyses were based on the responses of 2229 students who were enrolled in 93 class sections taught by 53 different instructors. For a description of these subjects see Table 1.3.

Analytic Methods. Analysis of variance procedures were used to assess the effects of class size and meeting time on SOST ratings. All analyses were performed on class means (N=93). For a further description of analytic methods see section 3.6.

3.71 The Effect of Class Size on SOST Ratings

Each of the 93 class sections was categorized as either "small", "medium" or "large". Operationally, a small class was defined as having fewer than 20 students; a medium class as having 20 to 50 students, and a large class as having more than 50 students. The mean class size for all sections was approximately 24 (23.97).

The results of the analyses of variance showed significant differences for 10 SOST items when section mean responses were classified by class size (Table 3.14). Mean section ratings generally favoured small and/or medium sized classes over large classes. These findings support previous work on the effects of class size on teaching evaluations (Table 2.10).

A long-standing debate among educators and psychologists has centred around the effect of class size on school learning. Do students

Table 3.14 EFFECT OF CLASS SIZE ON SOST RATINGS[1]
(N=93)

Item	Means and (Standard Deviations) by Class Size			
	Small (<20)	Medium (20-50)	Large (>50)	F ratio
9	1.72 (.31)	1.82 (.30)	2.01 (.65)	3.34*
10	2.12 (.53)	2.19 (.55)	2.27 (.62)	0.51(NS)
11	3.85 (.57)	3.81 (.48)	3.65 (.50)	1.14(NS)
12	1.96 (.59)	1.98 (.55)	1.79 (.36)	0.98(NS)
13	2.02 (.46)	1.76 (.37)	2.13 (.54)	4.88**
14	1.55 (.40)	1.48 (.44)	1.42 (.17)	0.97(NS)
15	1.67 (.37)	1.67 (.31)	1.96 (.45)	5.34**
16	3.72 (.50)	3.68 (.29)	3.56 (.37)	1.76(NS)
17	1.74 (.27)	1.82 (.32)	2.13 (.43)	10.45****
18	2.20 (.49)	2.27 (.44)	2.38 (.62)	0.84(NS)
19	3.34 (.51)	3.41 (.41)	3.17 (.49)	1.71(NS)
20	2.58 (.51)	2.74 (.39)	2.73 (.47)	1.30(NS)
21	2.19 (.42)	2.38 (.39)	2.51 (.48)	4.49*
22	2.27 (.47)	2.45 (.42)	2.66 (.34)	6.13**
23	2.46 (.64)	2.31 (.27)	2.58 (.38)	2.08(NS)
24	3.64 (.44)	3.75 (.31)	3.39 (.34)	6.04**
25	3.44 (.38)	3.46 (.29)	3.51 (.31)	0.37(NS)
26	3.41 (.46)	3.72 (.42)	3.50 (.62)	3.44*
27	3.01 (.50)	2.60 (.34)	2.78 (.32)	8.43***
28	2.14 (.34)	2.29 (.28)	2.50 (.31)	9.38***

[1] Of the 93 class sections 40 were classified as "small" (fewer than 20 students), 30 were classified as "medium" (20-50 students), and 23 were classified as "large" (more than 50 students). Average class size was 24.

actually "learn more" in a small class? While evidence exists both for and against this proposition, it is generally agreed that teachers and students alike prefer smaller classes to larger ones. The results of their analyses should, therefore, not be surprising.

3.72 The Effect of Class Meeting Time on SOST Ratings

Significant differences were found for 3 SOST items when section mean responses were classified by class meeting time (Table 3.15). These findings do not support the often-heard contention that morning classes are rated more favourably than mid-day and afternoon classes.

The results indicate that students enrolled in afternoon and evening classes feel that verbal and written feedback have been more constructive than students in other classes (Item 22). Furthermore, afternoon and evening students feel that the work required by the course was less intensive (Item 26) and they are less likely to indicate that the material was beyond their previous academic experience (Item 27).

IV SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

This section summarizes the findings of the study, presents the conclusions, and forwards several recommendations concerning the development and use of teaching evaluation instruments at the University of Windsor.

4.1 SUMMARY

4.11 Internal Consistency

The internal consistency of the SOST (using Cronbach's alpha coefficient) was found to be moderate to relatively high on three of the subscales (Sections A, B, and C). However, the alpha coefficients for Section D (.37) and Section E (.19) were unacceptably low. This finding is consistent with the factor analysis which shows that items in Sections D and E load on separate factors.

4.12 Stability

The stability coefficients for the SOST were found to be moderate to low, but generally within the range reported for other teaching

Table 3.15 EFFECT OF CLASS MEETING TIME ON SOST RATINGS[1]

(N=93)

Item	Means and (Standard Deviations) by Meeting Time			
	Morning	Mid-day	Afternoon/Evening	F ratio[2]
9	1.83 (.43)	1.84 (.44)	1.78 (.42)	0.13(NS)
10	2.23 (.61)	2.16 (.55)	2.15 (.52)	0.17(NS)
11	3.78 (.52)	3.80 (.49)	3.76 (.61)	0.05(NS)
12	1.91 (.48)	1.92 (.59)	1.95 (.50)	0.05(NS)
13	1.90 (.37)	1.95 (.53)	2.06 (.47)	0.76(NS)
14	1.51 (.30)	1.47 (.39)	1.52 (.40)	0.16(NS)
15	1.83 (.48)	1.65 (.35)	1.80 (.32)	2.11(NS)
16	3.64 (.46)	3.73 (.39)	3.67 (.41)	0.39(NS)
17	1.98 (.50)	1.78 (.27)	1.88 (.31)	2.70(NS)
18	2.33 (.61)	2.25 (.46)	2.23 (.48)	0.28(NS)
19	3.19 (.58)	3.44 (.39)	3.24 (.48)	2.73(NS)
20	2.79 (.53)	2.61 (.42)	2.63 (.47)	1.23(NS)
21	2.44 (.53)	2.31 (.40)	2.23 (.42)	1.46(NS)
22	2.61 (.45)	2.40 (.43)	2.25 (.39)	4.61*
23	2.42 (.36)	2.44 (.57)	2.47 (.49)	0.06(NS)
24	3.58 (.35)	3.62 (.45)	3.64 (.37)	0.16(NS)
25	3.42 (.31)	3.46 (.37)	3.52 (.29)	0.58(NS)
26	3.71 (.36)	3.53 (.52)	3.34 (.58)	3.42*
27	2.63 (.30)	2.87 (.52)	2.94 (.38)	3.76*
28	2.35 (.32)	2.24 (.35)	2.27 (.35)	0.84(NS)

[1] Twenty-seven (27) classes met in the morning (9:00 or 10:00 A.M.), 43 classes met at mid-day (11:00, 12:00 or 1:00) and 23 classes met in the afternoon or evening (2:00 - 7:00 P.M.).

[2] *$p < .05$

evaluation instruments. Mean coefficients were: .58 (7 day interval), .53 (14 day interval), .49 (21 day interval) and .43 (28 day interval).

4.13 Relationship Between Ratings and Student Achievement

Low but significant correlations were found between 11 of the SOST items and student achievement in an introductory psychology course. In all cases, the significant correlations indicated a positive relationship between student ratings and achievement.

These findings are taken as evidence that the instrument possesses a certain degree of criterion-related validity for it argues that instructors who receive favourable ratings are more successful in facilitating learning among their students than instructors who receive less favourable ratings.

4.14 Factor Analysis

Five factors emerged from the factor analysis procedure. These factors accounted for approximately 55% of the variance in the item responses. In general, factor loadings were moderate to low and the interpretation of the factor structure was complicated by a significant number of cross-loadings.

The five factors were identified as follows:

Factor I - Instructional Skill (Items 9, 10, 11, 12, 18, 19, 20, and 21). This is a general factor which measures the instructors ability to communicate with and motivate students. The large number of items indicates a possible "halo effect."

Factor II - Interaction (Items 17, 20, 21, 22, and 28). This factor is a measure of rapport and the general level of verbal and written exchanges between students and the instructor.

Factor III - Workload (Items 25 and 26). This factor is a measure of the amount of work required in the course.

Factor IV - Organization (Items 12, 13, and 14). This factor is a general assessment of the instructor's preparedness and clarity in explaining course objectives and

requirements.

Factor V - Feedback (Item 24). This factor measures the extent to which the student is able to assess his progress and achievement in the course.

4.15 Effect of Student Variables on Ratings

The results of a series of analyses of variance indicated that:

a) the student's major (faculty affiliation) may affect his evaluations of courses and instructors

b) upper-level and graduate-level students tend to rate instructors more favourably than lower-level students

c) students who feel that their performance is "superior" or "above average" relative to others in the class tend to give their instructors better ratings.

d) elective courses are rated more favourably than required courses

e) students who report that their effort in the course was "excellent" or "above average" relative to their effort in other courses rate the instructor and the course more favourably than other students.

4.16 Effect of Instructor Variables on Ratings

Analyses of variance indicated that, in several cases (5 items), senior faculty members (Professor; Associate Professor) are evaluated more favourably than junior faculty members (Assistant Professor; others). Furthermore, male instructors receive more positive ratings than female instructors on 7 items.

4.17 Effect of Class Variables on Ratings

A final series of analyses of variance showed that small and medium-sized classes tend to receive more favourable ratings than large classes (10 items) but that class meeting time generally has no effect on student evaluations.

4.2 CONCLUSIONS AND RECOMMENDATIONS

Although the SOST seems to possess some relatively positive psychometric qualities, namely criterion-related validity and reasonable stability, the instrument should not be adopted in its present form without revision and further testing. Of particular concern is the factor structure and the associated internal consistency as well as student, instructor, and class variables which, in some cases, consistently affect student ratings.

Specifically, the existing subscale organization of the SOST does not accurately reflect the factor structure of the instrument. This is shown by the magnitude of loadings within factors and is manifest in low internal consistency coefficients among 2 of the 5 subscales.

Furthermore, serious consideration must be given to factors which affect student ratings, especially those over which instructors have little or no control (ex: class size, required/elective course, instructor sex and rank). If major decisions concerning faculty fate are to be based, in part, on student evaluations; then ratings must be adjusted to take these factors into account.

One of the avowed purposes of student evaluations is to provide feedback to instructors who wish to improve their teaching effectiveness. Perhaps this is the most important use of these ratings. Unfortunately, there is considerable doubt whether items stated in global terms (such as "The instructor motivated one to put forth a good effort") provide this feedback in sufficiently specific terms. Compare, for example, items on the SOST with the following items taken from Murray's (1977) Teacher Rating Form:

The instructor:

14. moves back and forth in front of class
17. asks students questions during lecture
20. addresses individual students by name
23. maintains eye contact with students
31. gestures with hands and arms while speaking

Items such as these which are based on specific, observable,

teaching behaviours are considerably more useful to instructors and are probably more reliable since students need not make inferences concerning the instructor's motivations or general abilities.

Based on these and other considerations the following recommendations are forwarded:

Recommendation 1: If the existing instrument is to be retained, the following revisions should be considered:

a) The existing subscale organization of the SOST should be dropped in favour of either random ordering of items without subscale headings or use of subscales which reflect the factor structure of the instrument (Instructional Skill, Interaction, Workload, Organization, and Feedback).

b) Items 15, 16, 23, and 27 should be omitted. These items have reasonably low factor loadings and therefore tend to obscure the significance of individual factors.

c) Items 12, 20, and 21 cross-load significantly on two factors. These items should be reworded or deleted.

d) The Likert scales associated with Items 25 and 26 should be reworded so that they are consistent with other items (ex: strongly agree, agree.....)

e) Other instruments should be examined to find additional and appropriate replacement items within factors (Appendix D)

f) An additional subscale containing items on grading procedures should be added.

Recommendation 2: Several other existing instruments should be reviewed as possible alternatives to the SOST (Appendix D). The following instruments were administered to students who were enrolled in the second semester of an introductory psychology course (Psychology 115b):

a) Murray's Teacher Rating Form (1977)

b) Educational Testing Service's Student Instructional Report (1975)

c) Kansas State University's IDEA Survey Form (1975)

d) Purdue University's "Cafeteria" Instructional Rating Form (1975)

Correlations between items on each of these instruments and SOST items are given in Appendix D. In addition to the instruments listed above, Frey's Instructional Rating Form (Frey, 1973 and Appendix D) should be given serious consideration.

The advantages of these instruments over the SOST are several. As was mentioned earlier, Murray's instrument is based on specific, observable behaviours and therefore may provide more informative feedback to instructors. It may also be more reliable.

The advantage of the ETS, Kansas State and Purdue instruments is that a considerable amount of normative data is already available based on classes in a wide range of academic disciplines, class sizes, instructor ranks and so forth. Use of these instruments and their normative scales would reduce the problem of adjusting for differences in student, instructor and class variables.

Finally, items of Frey's Instructional Rating Form have been shown to correlate very well with student achievement ($r \sim .90$). This implies validities considerably better than the SOST. In addition, factor loadings on each of Frey's 7 subscales are consistently high (approximately .90).

Recommendation 3: Further studies should re-examine the effects of student, instructor, and class variables on ratings. Before any rating system is institutionalized, a method of adjusting ratings for these variables must be developed.

Recommendation 4: Further studies should examine differences among ratings in the various departments, schools, and faculties of the University to determine how best to adjust for differences in academic disciplines and instructional styles.

BIBLIOGRAPHY

Aleamoni, L.M., and Graham, M.H. "The Relationship Between CEQ Rating and Instructor's Rank, Class Size and Course Level." Journal of Educational Measurement 11(2), (Fall 1975): 189-202.

Aleamoni, L.M., and Spencer, R.E. 1973. "The Illinois Course Evaluation Questionnaire: A description of its development and a report of some of its results." Educational and Psychological Measurement 33(3): 669-684.

Aleamoni, L. M., and Yimer, M. 1973. "An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness." Journal of Educational Psychology 64: 274-277.

American Psychological Association Standards for Educational and Psychological Tests. Washington, D.C., APA, 1974.

Anastasi, Anne Psychological Testing, 4th ed. New York, MacMillan Publishing Co., 1976.

Bausell, R.B., Schwartz, Stanley, and Purohit, Anal. "An Examination of the Conditions Under which Various Student Rating Parameters Replicate Across Time." Journal of Educational Measurement 12 (Winter 1975): 273-280.

Bendig, A.W. "A Preliminary study of the effect of academic level, sex, and course variables on student rating of psychology instructors." Journal of Psychology 34 (1952): 2-126.

Bendig, A.W. "A factor analysis of student ratings of psychology instructors on the Purdue Scale." Journal of Educational Psychology (1954): 45: 385-393.

Bousefield, W.A. "Students' ratings of qualities considered desirable in college professors." School and Society 1940, 51: 253-256.

Bresler, J.B. "Teaching effectiveness and government awards." Science, 1968, 169: 164-167.

Brown, William. "Some experimental results in the correlation of mental abilities." British Journal of Psychology 3, 296-322.

Butts, David P. and Wm. R. Capie, "Evaluating teachers using teacher performance." Paper presented at the 50th Annual meeting of the National Association for Research in Science Teaching. Cincinnati, 1977.

Centra, J.A. "Self-ratings of college teachers: a comparison with student ratings". Journal of Educational Measurement 10(4), (Winter 1973): 287-295.

Centra, J. 1974. "College teaching: Who should evaluate it?" Princeton (N.J.): Educational Testing Service.

Centra, John A. "Colleagues as Raters of Classroom Instruction." Journal of Higher Education 1975, 46(3): 327-337.

Clinton, R.J. "Qualities college students desire in college instructors." School and Society, 1930, 32: 702.

Cohen, J., & Humphreys, L.G. Memorandum to faculty, University of Illinois, Department of Psychology, 1960 (mimeographed).

Costin, F. "Intercorrelations between students' and course chairmen's ratings of instructors." University of Illinois, Division of General Studies, 1966.

Costin, F. "A Graduate Course in the Teaching of Psychology: Description and Evaluation." Journal of Teacher Education, 19 (1968): 425-432.

Costin, F., Greenough, W.T., & Menger, R.J. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness." Review of Educational Research, 41 (1971), 511-535.

Creager, J.A. "A multiple-factor analysis of the Purdue Rating Scale for Instructors" Purdue University Studies in Higher Education, 1950, 70: 75-96.

Crittenden, Kathleen S., Norr, James L., & LeBailly, Robert K. "Size of University Classes and Student Evaluation of Teaching." Journal of Higher Education 1975, 46,(4): 461-470.

Cronbach, Lee J. Essentials of Psychological Testing. New York, N.Y.: Harper & Row, 1970.

Deshpande, A.S., Webb, S.C., & Marks, E. "Student perceptions of engineering instructors behaviours and their relationships to the evaluation of instructors and courses." American Educational Research Journal. 1970, 7: 289-305.

Downie, N.M. "Student evaluation of faculty." Journal of Higher Education 23, (1952): 495-496.

Doyle, K.O. Jr. 1972. "Construction and evaluation of scales for rating Instructors." Dissertation Abstracts International, 1972, 33 (5-A), 2163.

Doyle, Kenneth O. Jr. Student Evaluation of Instruction. Lexington, Massachusetts: D.C. Heath and Company, 1975.

Drucker, A.J., and Remmers, H.H. "Do alumni and students differ in their attitudes toward instructors?" Purdue University Studies in Higher Education, 1950, 70: 62-64.

Educational Testing Service. Student Instructional Report Comparative Data Guide 1975-1976. ETS College and University Programs, Box 2813, Princeton, New Jersey, 1975.

Elliott, D.H. "Characteristics and relationships of various criteria of college and university teaching. Purdue University Studies in Higher Education, 1950, 70: 5-61.

Elsmore, Patricia B., and Lapointe, Karen A. "Effects of Teacher Sex and Student Sex on the Evaluation of College Instructors" Journal of Educational Psychology, 66, (1974): 386-389.

French, G.M. "College students' concept of effective teaching determined by an analysis of teacher ratings." Dissertation Abstracts, 1957, 17: 1380-1381.

Frey, P.W. "Comparative judgement scaling of student course ratings." American Educational Research Journal, 1973, 10: 149-154.

Frey, P.W. 1973. "Student ratings of teaching: Validity of several rating factors." Science, 182: 83-85.

Gadzella, B.M. "College student views and ratings of an ideal professor." College and University, 1968, 44: 89-96.

Gage, N.L. "The appraisal of college teaching: An analysis of ends and means." Journal of Higher Education, 1961, 32: 17-22.

Gessner, P.K. "Evaluation of instruction". Science 1973, 180: 566-570.

Gibb, C.E. "Classroom behavior of the college teacher." Educational and Psychological Measurement, 1955, 15: 254-263.

Gulliksen, Harold. Theory of Mental Tests. New York, John Wiley & Sons, Inc., 1965.

Hartley, E.L., & Hogan, T.P. "Some additional factors in student evaluation of courses." American Educational Research Journal, 1972, 9: 241-250.

Harvey, J.N., and Barker, D.G. "Student Evaluation of Teaching Effectiveness." Improving College and University Teaching, 1970, 18: 275-278.

Hayes, J.R. "Research, teaching and faculty fate." Science 1971, 172: 227-230.

Hildebrand, M., Wilson, R.C., and Dienst, E.R. 1971. Evaluating university teaching. Berkeley: Center for Research and Development in Higher Education, University of California, Berkeley.

IDEA - The Instructional Development and Effectiveness Assessment System Center for Faculty Evaluation and Development in Higher Education, 1627 Anderson Avenue, Box 3000, Manhattan, Kansas, 1975.

Isaacson, R.L., McKeachie, W.J., Milholland, J.E., Lin, Y.G., Hofeller, M., Baerwaldt, J.W., & Zinn, K.L. Dimensions of student evaluations of teaching. Journal of Educational Psychology, 1964, 55: 344-351.

Kennedy, W.R. "Grades Expected and Grades Received - Their relationship to students' evaluation of faculty performance." Journal of of Educational Psychology, 67 (1) (Feb. 1975): 109-115.

Kohlan, R.G. "A comparison of faculty evaluations early and late in the course." Journal of Higher Education, 1973, 44: 22-25.

Kooker, E.W. 1968. "The relationship of college grades to course ratings on student selected items." The Journal of Psychology, 69: 209-215.

Kulik, James A., and McKeachie, Wilbert J. "The Evaluation of Teachers in Higher Education". In Review of Research in Education. Itasca, Illinois, F.E. Peacock Publishers, Inc. 1975.

Langen, T.D.F. "Student assessment of teaching effectiveness." Improving College and University Teaching, 1966, 14: 22-25

Lovell G.D. & Haner, C.F. "Forced-choice applied to college faculty rating." Education and Psychological measurement, 1955; 15: 291-304.

Maslow, A.H., and Zimmerman, W. 1956. "College teaching ability, scholarly activity and personality." Journal of Educational Psychology, 47: 185-189.

McDaniel, E.D., and Feldhusen, J.F. "Relationships between faculty ratings and indexes of service and scholarship." Proceedings, 78th Annual Convention APA, 1970, 619-620.

McDaniel, E.D., & Feldhusen, J.F. "College teaching effectiveness." Today's Education, 1971, 60: 27.

McKeachie, W.J., Isaacson, R., and Milholland J. 1964. "Research on the characteristics of effective college teaching." Final report Cooperative Research Project No. OE850, Office of Education, Department of Health, Education and Welfare, Ann Arbor.

McKeachie, W.J., & Lin, Y.G. Multiple discriminant analysis of student ratings of college teachers. Unpublished manuscript, University of Michigan, 1973.

McKeachie, W.J., Lin, Y., and Mann, W. "Student ratings of teacher effectiveness: Validity studies." American Educational Research Journal, 1971, 8: 435-445.

Miller, R.I. Developing Programs for Faculty Evaluation. San Francisco, California, Jossey-Bass, Inc., 1974.

Miller, R.I. Evaluating Faculty Performance San Francisco: Jossey-Bass, 1972.

Marsh, H.W., Fleiner, H., and Thomas, C.S. "Validity and Usefulness of Student Evaluations of Instructional Quality." Journal of Educational Psychology 67 (6), (1975): 833-839.

Morsh, J.E., & Wilder, E.W. "Identifying the effective instructor: A review of the quantitative studies, 1900-1952. (Research Bull. ARPTRC-TR-54-44). Air Force Personnel and Training Research Center, 1954.

Morsh, J.E., Burgess, G.G. and Smith, P.N. "Student achievement as a measure of instructor effectiveness." Journal of Educational Psychology, 1956, 47: 79-88.

Murray, Harry G. A Guide to Teaching Evaluation. Toronto, Ontario Confederation of University Faculty Associations, 1973.

Murray, Harry G. "Lecturing-Classroom Behaviours of Social Science Lecturers Receiving Low, Medium and High Teacher Ratings." OUPID Newsletter 14 (February 1977): 3-5.

Perry, R.R. Evaluation of teaching behaviour seeks to measure effectiveness. College and University Business, 1969, 47: 18-22.

Pogue, F.R., Jr. "'Students' Ratings of the 'Ideal Teacher'". Improving College and University Teaching, 15 (1967): 133-136.

Pohlmann, John T. "A Description of Teaching Effectiveness as Measured by Student Ratings." Journal of Educational Measurement, 12 (Spring 1975): 49-54.

Purdue University "Cafeteria" Instructional Rating Form. Purdue Research Foundation, West Lafayette, Indiana, 1975.

Quereshi, M.Y. "Teaching Effectiveness and Research Productivity." Science, 1968, 161: 1160.

Rayder, N.F. "College student ratings of instructors." Journal of Experimental Education, 1968, 37: 76-81.

Remmers, H.H. "Appraisal of college teaching through ratings and student opinion. "In 27th Yearbook of the National Society of College Teachers of Education. Chicago: University of Chicago Press, 1939.

Remmers, H.H. Manual of instructions for the Purdue Rating Scale for Instructors, (Rev. ed.) West Lafayette, Ind.: University Book Store, 1960.

Remmers, H.H., and Brandenburg, G.C. 1927. "Experimental data on the Purdue Rating Scale for Instructors." Educational Administration and Supervision, 13: 519-527.

Remmers, R.H., & Elliott, D.N. "The Indiana College and University Staff-Evaluation Program." School and Society, 1949, 70: 168-171.

Remmers, H.H., Shock, N.W., and Kelley, E.L. "An empirical study of the Spearman-Brown Formula as applied to the Purdue Rating Scale." Journal of Educational Psychology, 1927, 18: 187-195.

Remmers, H.H. & Weisbrodt, J.A. Manual of Instructions for the Purdue Rating Scale for Instructions, West Lafayette, Indiana: University Book Store, 1965.

Rodin, M. and Rodin, B. 1972. "Student evaluations of teachers." Science 1977: 1164-1166.

Root, A.R. 1931. "Student Ratings of Teachers. Journal of Higher Education, 2, 311-315.

Skanes, G.R., and Sullivan, A.M. "Validity of Student Evaluation of Teaching and the Character of Successful Instructors." Journal of Educational Psychology, 66 (4), 584-590, 1974.

Smalzreid, N.T., & Remmers, H.H. "A factor analysis of the Purdue Rating Scale for Instructors. "Journal of Educational Psychology, 1943, 34: 363-367.

Solomon, D. "Teacher behaviour dimensions, course characteristics, and student evaluations of teachers," American Educational Research Journal, 1966, 3: 35-47.

Spearman, Charles. "Correlation calculated with faculty data." British Journal of Psychology, 3, 271-295.

Student Evaluations Committee. "Report of the Senate Student Evaluations Committee." Faculty Senate, University of Windsor, December 16, 1975.

Thorndike, R.L. Personnel Selection. New York: Wiley, 1949.

Turner, R.L. "Good teaching and its contexts." Phi Delta Kappan, 1970, 51: 155-158.

Voeks, V.W. "Publications and teaching effectiveness. "Journal of Higher Education, 1962, 33: 212.

Voeks, V.W., & French, G.M. "Are student ratings of teachers affected by grades?" Journal of Higher Education, 1960, 31: 330-334.

Wherry, R.J. 1952. "Control of bias in ratings." Department of the Army, The Adjutant General's Office, Personnel Research and Procedures Division, Personnel Research Branch. PRS Reports 914, 915, 919, 920 and 921.

Winer, B.J., Statistical Principles in Experimental Design. New York, N.Y.: McGraw-Hill Book Company, 1962, 1971.

Wood, K., Lensky, A.S., and Strauss, M.A. "Class Size and Student Evaluation of Faculty." Journal of Higher Education, 1974, 45: 542-534.

APPENDIX A

THE STUDENT OPINION SURVEY OF TEACHING (SOST)

AND

INTERCORRELATIONS BASED ON 2229 STUDENT RESPONSES

STUDENT OPINION SURVEY OF TEACHING

PART I General Information

1. My major is in: Arts (A) Social Science (B) Science & Math (C) Business (D) Other (E)

2. This course is part of my honours/general program. (A) (B)

3. I have completed the following number of University level full courses (two half courses equal one full): 0--2 (A) 3--7 (B) 8--12 (C) 13--17 (D) 18--. (E)

4. Rating myself against the performance of other students in the class, I see myself in one of the following groups: superior (A), above average (B), average (C), below average (D), failing (E).

5. This course was compulsory. YES (A) NO (B) NOT SURE (C)

6. My attendance and punctuality have been consistently good. YES (A) NO (B)

7. Compared to other courses I have taken, I consider my effort in this course to have been: excellent (A), above average (B), average (C), below average (D), poor (E).

8. I have found the material in this course to be inherently difficult. YES (A) NO (B)

PART II ALL FOLLOWING QUESTIONS ARE RATED ON A FIVE-POINT SCALE FROM STRONGLY AGREE TO STRONGLY DISAGREE EXCEPT WHERE NOTED.

Section A. Communication (Instructor - Group Interaction)

9. The instructor is clear and audible.

Strongly Agree (A) Agree (B) Not Sure (C) Disagree (D) Strongly Disagree (E)

10. The instructor presented material in a coherent manner, emphasizing major points and making relationships clear.

Strongly Agree (A) Agree (B) Not Sure (C) Disagree (D) Strongly Disagree (E)

11. Course material was disorganized and hindered understanding.

Strongly Agree (A) Agree (B) Not Sure (C) Disagree (D) Strongly Disagree (E)

12. The instructor was consistently prepared for class.

Strongly Agree (A) Agree (B) Not Sure (C) Disagree (D) Strongly Disagree (E)

13. The instructor was clear on what was expected regarding course requirements, assignments, exams, etc.

Strongly Agree (A) Agree (B) Not Sure (C) Disagree (D) Strongly Disagree (E)

14. The instructor's attendance and punctuality have been consistently good.

Strongly Agree (A) Agree (B) Not Sure (C) Disagree (D) Strongly Disagree (E)

Section B. Communication (Instructor - Individual Interaction)

15. The instructor encouraged and readily responded to student questions.

Strongly Agree (A) Agree (B) Not Sure (C) Disagree (D) Strongly Disagree (E)

16. The instructor has not been readily available for consultation by appointment or otherwise.

Strongly Agree (A) Agree (B) Not Sure (C) Disagree (D) Strongly Disagree (E)

17. The instructor maintained a generally helpful attitude toward students and their problems.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

Section C. Motivation and Impact

18. The instructor made this course as interesting as the subject matter would allow.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

19. The instructor did not increase my interest in the subject matter of the course.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

20. The instructor motivated me to put forth a good effort.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

21. The instructor was successful in making difficult material understandable.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

Section D. Feedback

22. Verbal or written comments on assignments have been constructive.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

23. The evaluation system for this course was fairly applied.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

24. Throughout this course, I have not been able to assess my progress and achievement.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

25. The instructor's expectations for student performance were very low (A), low (B), average (C), high (D), very high (E).

Section E. Standards

26. The amount of work required for this course has been very light (A), light (B), average (C), heavy (D), very heavy (E).

27. The material covered in this course has been beyond my previous academic experience.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

28. The assignments provided a valuable learning experience.

Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
A	B	C	D	E

Has this questionnaire given you an adequate opportunity to express your opinion about the instruction in this course?

YES	NO
A	B

CORRELATION COEFFICIENTS..

	ITEM09	ITEM10	ITEM11	ITEM12	ITEM13	ITEM14	ITEM15	ITEM16	ITEM17	ITEM18
ITEM09	1.00000	0.58654	-0.36462	0.38735	0.36446	0.18988	0.36565	-0.16992	0.34692	0.48596
ITEM10	0.58654	1.00000	-0.50892	0.48868	0.40607	0.23827	0.40777	-0.20757	0.38976	0.54994
ITEM11	-0.36462	-0.50892	1.00000	-0.36140	-0.34142	-0.19066	-0.26587	0.18867	-0.24055	-0.40073
ITEM12	0.38735	0.48868	-0.36140	1.00000	0.40496	0.37890	0.26283	-0.19678	0.27898	0.41456
ITEM13	0.36446	0.40607	-0.34142	0.40496	1.00000	0.26102	0.33714	-0.20676	0.34909	0.37458
ITEM14	0.18988	0.23827	-0.19066	0.37890	0.26102	1.00000	0.28914	-0.15611	0.25929	0.21006
ITEM15	0.36565	0.40777	-0.26587	0.26283	0.33714	0.28914	1.00000	-0.26521	0.56455	0.43347
ITEM16	-0.16992	-0.20757	0.18867	-0.19678	-0.20676	-0.15611	-0.26521	1.00000	-0.33770	-0.20610
ITEM17	0.34692	0.38976	-0.24055	0.27898	0.34909	0.25929	0.56455	-0.33770	1.00000	0.46577
ITEM18	0.48596	0.54994	-0.40073	0.41456	0.37458	0.21006	0.43347	-0.20610	0.46577	1.00000
ITEM19	-0.35958	-0.41737	0.33177	-0.26591	-0.29261	-0.14684	-0.34368	0.27407	-0.34615	-0.50940
ITEM20	0.34002	0.41711	-0.25470	0.32117	0.30546	0.21659	0.37769	-0.22478	0.42659	0.50653
ITEM21	0.45396	0.59953	-0.42550	0.35905	0.35861	0.21015	0.40331	-0.23080	0.42529	0.54413
ITEM22	0.20870	0.24377	-0.15196	0.15991	0.21521	0.15920	0.26765	-0.20591	0.32075	0.32775
ITEM23	0.19613	0.25106	-0.22017	0.19531	0.30237	0.09693	0.19163	-0.16436	0.22100	0.23791
ITEM24	-0.21180	-0.20933	0.27600	-0.15531	-0.25266	-0.06854	-0.17168	0.16316	-0.15212	-0.18428
ITEM25	-0.03016	-0.04321	-0.00801	-0.00777	0.02166	-0.04135	-0.06972	0.07150	-0.00562	-0.03109
ITEM26	0.04321	0.04086	-0.05410	0.04573	0.02574	-0.01354	-0.01128	0.03791	0.01701	0.05731
ITEM27	0.00687	-0.02943	0.03954	0.02109	-0.02574	0.04876	-0.00167	-0.01522	0.08053	0.01328
ITEM28	0.25351	0.26370	-0.22558	0.19081	0.25167	0.14246	0.23989	-0.20103	0.29705	0.33263

	ITEM19	ITEM20	ITEM21	ITEM22	ITEM23	ITEM24	ITEM25	ITEM26	ITEM27	ITEM28
ITEM09	-0.35958	0.34002	0.45396	0.20870	0.19613	-0.21180	-0.03016	0.04321	0.00687	0.25351
ITEM10	-0.41737	0.41711	0.59953	0.24377	0.25106	-0.20933	-0.04321	0.04086	-0.02943	0.26370
ITEM11	0.33177	-0.25470	-0.42550	-0.15196	-0.22017	0.27600	-0.00801	-0.05410	0.03954	-0.22558
ITEM12	-0.26591	0.32117	0.35905	0.15991	0.19531	-0.15531	-0.00777	0.04573	0.02109	0.19081
ITEM13	-0.29261	0.30546	0.35861	0.21521	0.30237	-0.25266	0.02166	0.02574	-0.02574	0.25167
ITEM14	-0.14684	0.21659	0.21015	0.15920	0.09693	-0.06854	-0.04135	-0.01354	0.04876	0.14246
ITEM15	-0.34368	0.37789	0.40331	0.26765	0.19163	-0.17168	-0.06972	-0.01128	-0.00167	0.23989
ITEM16	0.27407	-0.22478	-0.23080	-0.20591	-0.16436	0.16316	0.07150	0.03791	-0.01522	-0.20103
ITEM17	-0.34619	0.42659	0.42529	0.32975	0.22100	-0.15212	-0.00562	0.01701	0.08053	0.29705
ITEM18	-0.50940	0.50653	0.54413	0.27477	0.24791	-0.18428	-0.03109	0.05731	0.01328	0.33263
ITEM19	1.00000	-0.47904	-0.42846	-0.21044	-0.18669	0.29555	0.13512	0.06604	-0.02204	-0.29501
ITEM20	-0.47904	1.00000	0.51186	0.35790	0.26331	-0.12892	-0.07544	-0.03118	0.03584	0.40665
ITEM21	-0.42846	0.51186	1.00000	0.33790	0.29923	-0.17578	-0.02704	0.02531	0.09517	0.35578
ITEM22	-0.21044	0.35790	0.33790	1.00000	0.30490	-0.10972	0.03588	0.00535	0.02239	0.27345
ITEM23	-0.18669	0.26331	0.29923	0.30490	1.00000	-0.23155	0.09517	0.15135	-0.02541	0.27335
ITEM24	0.29555	-0.12892	-0.17578	-0.10972	-0.23155	1.00000	-0.07206	0.03791	0.15330	-0.14246
ITEM25	0.13512	-0.07544	-0.02204	0.03584	0.09517	-0.07206	1.00000	0.29384	-0.03435	0.23989

	ITEM19	ITEM20	ITEM21	ITEM22	ITEM23	ITEM24	ITEM25	ITEM26	ITEM27	ITEM28
ITEM26	-0.06606	-0.03131	0.02516	0.00535	0.15135	0.03791	0.29384	1.00000	-0.16412	0.14465
ITEM27	0.07210	0.02631	-0.01283	0.02299	-0.02541	0.15330	-0.03435	-0.16412	1.00000	0.07730
ITEM28	-0.29501	0.40665	0.35578	0.27290	0.27335	-0.14246	0.23989	0.14465	0.07730	1.00000

80

APPENDIX B

FOLLOW-UP LETTER

University of Windsor

Dear Prof.

February 17, 1977.

Thank you for agreeing to cooperate in the validation of the Student Opinion Survey of Teaching. As you probably know, this instrument was developed by the Faculty Senate Student Evaluations Committee and is being considered for university-wide adoption. Dr. David Reynolds and I have received an O.U.P.I.D. Grant to examine the reliability and validity of the SOST and to recommend changes or revisions. A copy of the instrument is attached.

So that we might collect data without unnecessarily disrupting your class schedule, I'd like to ask you to complete the table shown below by indicating:

(1) the course name/number
(2) date you wish the evaluation
(3) class meeting time
(4) meeting place (building and room #)
(5) whether the evaluator (research assistant) should distribute the instrument at the beginning or at the end of the period
(6) approximate enrollment

Depending on class size, the evaluation requires approximately 10-15 minutes of class time.

Course	Evaluation Date	Meeting Time	Meeting Place	Distribute at Beginning/end of period	Approximate Enrollment

All data will be treated confidentially. However, if you wish, a complete printout of your own evaluation can be provided. Please indicate if you wish to receive a copy of your evaluation (_____yes; _____ no).

Thank you again for your cooperation and please return this to me by Friday, February 25th.

Joel J. Mintzes
Assistant Professor
Department of Biology

APPENDIX C

"NORMATIVE" DATA BASED ON 93 CLASSES

"NORMATIVE" DATA FOR SOST[1]

(N=93 CLASSES)

Item	Mean	Standard Deviation of Means	Average Standard Deviation
9	1.82	.43	.72
10	2.18	.56	.83
11	3.79	.53	.88
12	1.93	.53	.73
13	1.96	.47	.83
14	1.49	.37	.55
15	1.74	.39	.71
16	3.69	.41	.84
17	1.86	.37	.70
18	2.27	.51	.87
19	3.32	.48	.99
20	2.67	.47	.91
21	2.33	.44	.80
22	2.43	.44	.89
23	2.44	.49	.99
24	3.62	.40	.96
25	3.46	.34	.78
26	3.53	.51	.81
27	2.82	.45	1.19
28	2.28	.34	.88

[1] Mean - this column contains the means of class report means. Means from combined reports and also those from individual class reports may be compared to these means.

Standard Deviation of the Means - this column contains the standard deviations of the class report means. It is appropriate to compare the standard deviations from combined reports with these figures.

Average Standard Deviation - this column contains the average standard deviations of the class report means. It is appropriate to compare the standard deviations in an individual instructor's class report with these figures.

APPENDIX D,

Pages 85-119,

REMOVED AT THE AUTHOR'S REQUEST.